DISSERTATION

SEMIPARAMETRIC REGRESSION IN THE PRESENCE OF COMPLEX VARIANCE STRUCTURES ARISING FROM SMALL ANGLE X-RAY SCATTERING DATA

Submitted by

Bruce D. Bugbee, Jr.

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2014

Doctoral Committee:

Advisor: F. Jay Breidt
Co-Advisor: Don Estep

Mary Meyer
Jennifer Hoeting
Karolin Luger

ABSTRACT

SEMIPARAMETRIC REGRESSION IN THE PRESENCE OF COMPLEX VARIANCE STRUCTURES ARISING FROM SMALL ANGLE X-RAY SCATTERING DATA

An ongoing problem in structural biology is how best to infer structural information for complex, biological macromolecules from indirect observational data. Molecular shape dictates functionality but is not always directly observable. There exists a wide class of experimental methods whose data can be used for indirectly inferring molecular shape features with varying degrees of resolution. Of these methods, small angle X-ray scattering (SAXS) is desirable due to low requirements on the sample of interest. However, SAXS data suffers numerous statistical problems that require the development of novel methodologies. A primary concern is the impact of radially reducing two-dimensional sensor data to a series of smooth mean and variance curves. Additionally, pronounced heteroskedasticity is often observed near sensor boundaries. The work presented here focuses on developing general model frameworks and implementation methods appropriate for SAXS data.

Semiparametric regression refers to models that combine known parametric structures with flexible nonparametric components. Three semiparametric regression model frameworks that are well-suited for handling smooth data are presented. The first model introduced is the standard semiparametric regression model, described as a mixed model with low rank penalized splines as random effects. The second model extends the first to the case of heteroskedastic errors, which violate standard model assumptions. The latent variance function in the model is estimated through an additional semiparametric regression, allowing for appropriate uncertainty estimation at the mean level. The final model considers a data structure unique to SAXS experiments. This model incorporates both radial mean and radial

variance data in hopes to better infer three-dimensional shape properties and understand experimental effects by including all available data.

Each of the three model frameworks is structured hierarchically. Bayesian inference is appealing in this context, as it provides efficient and generalized modeling frameworks in a unified way. The main statistical contributions of this thesis are from the specific methods developed to address the computational challenges of Bayesian inference for these models. The contributions include new Markov Chain Monte Carlo (MCMC) procedures for numerical approximation of posterior distributions and novel variational approximations that are extremely fast and accurate. For the heteroskedastic semiparametric case, known form posterior conditionals are available for all model parameters save for the regression coefficients controlling the latent model variance function. A novel implementation of a multivariate delayed rejection adaptive Metropolis (DRAM) procedure is used to sample from this posterior conditional distribution. The joint model for radial mean and radial variance data is shown to be of comparable structure to the heteroskedastic case and the new DRAM methodology is extended to handle this case. Simulation studies of all three methods are provided, showing that these models provide accurate fits of observed data and latent variance functions.

The demands of scientific data processing in the context of SAXS, where large data sets are rapidly attained, lead to consideration of fast approximations as alternatives to MCMC. Variational approximations or Variational Bayes describes a class of approximation methods where the posterior distribution of the parameters is approximated by minimizing the Kullback-Leibler divergence between the true posterior and a class of distributions under mild structural constraints. Variational approximations have been shown to be good approximations of true posteriors in many cases. A novel variational approximation for the general heteroskedastic semiparametric regression model is derived here. Simulation studies are provided demonstrating fit and coverage properties comparable to the DRAM results

at a fraction of the computational cost. A variational approximation for the joint model of radial mean and variance data is also provided but is shown to suffer from poor performance due to high correlation across a subset of regression parameters.

The heteroskedastic semiparametric regression framework has some strong structural relationships with a distinct, important problem: spatially adaptive smoothing. A noisy function with different amounts of smoothness over its domain may be systematically under-smoothed or over-smoothed if the smoothing is not spatially adaptive. A novel variational approximation is derived for the problem of spatially adaptive penalized spline regression, and shown to have excellent performance. This approximation method is shown to be able to fit highly oscillatory data while not requiring the traditional tuning and computational resources of standard MCMC implementations.

Potential scientific contribution of the statistical methodology developed here are illuminated with SAXS data examples. Analysis of SAXS data typically has two primary concerns: description of experimental effects and estimation of physical shape parameters. Formal statistical procedures for testing the effect of sample concentration and exposure time are presented as alternatives to current methods, in which data sets are evaluated subjectively and often combined in *ad hoc* ways. Additionally, estimation procedures for the scattering intensity at zero angle, known to be proportional to molecular weight, and the radius of gyration are described along with appropriate measures of uncertainty. Finally, a brief example of the joint radial mean and variance method is provided. Guidelines for extending the models presented here to more complex SAXS problems are also given.

ACKNOWLEDGEMENTS

I am grateful for the many opportunities that this work and my time at Colorado State University has presented. First and foremost, I owe the bulk of my success to my advisor Jay Breidt. His guidance and experience were crucial in my development as a researcher and statistician. Both in the classroom and outside of it, he is genuinely one of the best teachers I have ever had.

Throughout the course of my research I have worked closely with Mark van der Woerd from the Department of Molecular Biology and Biochemistry. His collaborations have been essential to my work. Additionally, his advice and assistance on career choices have been quite helpful. I owe special thanks to Jennifer Hoeting who, apart from serving on my committee, has also provided significant career and statistical guidance over my tenure at CSU. Additionally I would like to thank Don Estep, Mary Meyer, and Karolin Luger for serving on my committee.

This work has been partially supported by the Joint NSF/NIGMS Initiative to Support Research in the Area of Mathematical Biology (R01GM096192). The variational approximation for heteroskedastic semiparametric regression detailed in Chapter 4 has been submitted for publication in the *Journal of Computational Graphics and Statistics*. Preliminary work on jointly modeling radial mean and variance data (Sections 2.3 and 3.3) was awarded the Outstanding Student Poster award at the 2013 WNAR/IMS annual meeting and I would like to thank all those who participated for their feedback.

I am also thankful for the opportunity to participate in the Industrial Mathematical and Statistical Modeling Workshop for Graduate Students at the Statistical and Applied Mathematical Sciences Institute (SAMSI) during the summer of 2012. This workshop was a great

interdisciplinary experience and helped introduce topics that would become important to the work presented here, notably advanced MCMC methodologies.

I have greatly enjoyed the six years I have spent in Fort Collins and at CSU. This wonderful town has provided coutless opportunities for both personal and professional growth. I owe a great deal of my personal accomplishments over this time to Finnie and Tessa McMahon of McMahon Brazilian Jiu Jitsu for giving me a place to improve myself and find a second family away from my own. I have had the pleasure of sharing this experience with many amazing fellow students, particularly Wade Herndon, Grant Weller, and Stacy Edmondson. I would also like the thank the employees of Mugs Coffee Lounge for providing a caffeinated oasis where most of this dissertation was written.

Finally, I am most grateful for the love and encouragement provided my my parents Bruce and Kimm. They have stuck with me through this long journey and never failed to provide much-needed support. I thank them for instilling the determination and work ethic that got me to where I am today.

DEDICATION

For Mom and Dad. We did it.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

An ongoing problem in structural biology is how best to infer structural information for complex, biological macromolecules from indirect observational data. Molecular shape dictates functionality but is not always directly observable. There exists a wide class of experimental methods whose data can be used for indirectly inferring shape features with varying degrees of resolution. Of these methods, small angle X-ray scattering (SAXS) is desirable due to low requirements on the sample of interest. However, SAXS data suffers numerous statistical problems that require the development of novel methodologies. These problems include multiple sources of stochastic uncertainty as well as the application of *ad hoc* data processing procedures at the point of collection. A primary concern is the impact of radially reducing two-dimensional sensor data to a series of smooth mean and variance curves. Additionally, pronounced heteroskedasticity is often observed near sensor boundaries. The work presented here focuses on developing general model frameworks and implementation methods appropriate for SAXS data. The main statistical contributions of this thesis are from the specific methods developed to address the computational challenges of Bayesian inference for models appropriate for SAXS data. The contributions include new Markov Chain Monte Carlo (MCMC) procedures for numerical approximation of posterior distributions and novel variational approximations that are extremely fast and accurate. In this introductory chapter, Section 1.2 outlines the SAXS experimental procedure and data collection in detail. Section 1.3 details the background of treating semiparametric regression models as linear

mixed models through the use of low-rank penalized spline bases, as detailed in Ruppert et al. (2003).

Chapter 2 details the application of semiparametric regression models to SAXS data in the presence of complex variance structures. Semiparametric regression refers to models that combine known parametric structures with flexible nonparametric components. Three semiparametric regression model frameworks that are well-suited for handling smooth data are presented. The first model introduced is the standard semiparametric regression model, described as a mixed model with low rank penalized splines as random effects. The second model extends the first to the case of heteroskedastic errors, which violate standard model assumptions. The latent variance function in the model is estimated through an additional semiparametric regression, allowing for appropriate uncertainty estimation at the mean level. The final model considers a data structure unique to SAXS experiments. This model incorporates both radial mean and radial variance data in hopes to better infer three-dimensional shape properties and understand experimental effects by including all available data.

Chapter 3 introduces the computational procedures used to implement the models described in Chapter 2. Each of the three model frameworks has a hierarchical structure. Bayesian inference is appealing in this context, as it provides efficient and generalized modeling frameworks in a unified way. For the heteroskedastic semiparametric case, known form posterior conditionals are available for all model parameters save for the regression coefficients controlling the latent model variance function. A novel implementation of a multivariate delayed rejection adaptive Metropolis (DRAM) procedure is used to sample from this posterior conditional distribution. The joint model for radial mean and radial variance data is shown to be of comparable structure to the heteroskedastic case and the new DRAM methodology is extended to handle this case. Simulation studies of all three methods are pro-

vided, showing that these models provide accurate fits of observed data and latent variance functions.

Chapter 4 focuses on the demands of scientific data processing in the context of SAXS, where large data sets are rapidly, lead to consideration of fast approximations as alternatives to MCMC. Variational approximations or Variational Bayes describes a class of approximation methods where the posterior distribution of the parameters is approximated by minimizing the Kullback-Leibler divergence between the true posterior and a class of distributions under mild structural constraints. Variational approximations have been shown to be good approximations of true posteriors in many cases. A novel variational approximation for the general heteroskedastic semiparametric regression model is derived here. Simulation studies are provided demonstrating fit and coverage properties comparable to the DRAM results at a fraction of the computational cost. A variational approximation for the joint model of radial mean and variance data is also provided but is shown to suffer from poor performance due to high correlation across a subset of regression parameters.

In addition to work on variational approximation for models of SAXS data, Chapter 4 also details a variational approximation method for a distinct, important problem that has a strong structural relationship with heteroskedastic semiparametric regression: spatially adaptive smoothing. A noisy function with different amounts of smoothness over its domain may be systematically under-smoothed or over-smoothed if the smoothing is not spatially adaptive. A novel variational approximation is derived in Section 4.5 for the problem of spatially adaptive penalized spline regression, and shown to have excellent performance. This approximation method is shown to be able to fit highly oscillatory data while not requiring the traditional tuning and computational resources of standard MCMC implementations.

Chapter 5 expands on potential scientific contribution of the statistical methodology developed here with SAXS data examples. Analysis of SAXS data typically has two primary

concerns: description of experimental effects and estimation of physical shape parameters. Formal statistical procedures for testing the effect of sample concentration and exposure time are presented as alternatives to current methods, in which data sets are evaluated subjectively and often combined in *ad hoc* ways. Additionally, estimation procedures for the scattering intensity at zero angle, known to be proportional to molecular weight, and the radius of gyration are described along with appropriate measures of uncertainty. Finally, a brief example of the joint radial mean and variance method is provided. Guidelines for extending the models presented here to more complex SAXS problems are also given.

1.2 Introduction to Small Angle X-Ray Scattering

A motivating problem that is common in structural biology concerns the accurate identification of structural characteristics of macromolecules. Size and shape properties often are related to molecular properties and as such are highly interesting to scientists. One experimental method that is popular due to its applicability to a wide class of macromolecules and conditions is small angle X-ray scattering (SAXS). SAXS experiments provide low resolution structural information based on the bombardment of a sample of interest with X-rays and measurement of the resulting diffraction. Competing, higher-resolution imaging techniques often have physical constraints that make them ill-suited to complex or flexible macromolecules. For example, X-ray crystallography requires crystallization of the sample (**?**Bergfors, 1999). SAXS experiments are also relatively inexpensive to conduct. The major drawback to SAXS stems from the method's seeming inability to estimate high resolution characteristics. Physical inference of molecular structure is generally relegated to low-resolution characteristics, such as maximum linear dimension or radius of gyration (Glatter and Kratky, 1982). Compounding the resolution problem, experimental conditions, simplifying assumptions, and "black box" data processing are persistent issues surrounding SAXS data.

Figure 1.1 outlines the SAXS experimental procedure. First, a monochromatic beam of X-rays is fired at a sample of the molecule of interest in solution. The X-rays hit the sample and are scattered due to interactions with the molecules. These scattered X-rays are recorded on the detector mechanism, which usually consists of a phosphorescent screen connected to a CCD sensor (akin to the sensor in a digital camera) via fiber optics. The phosphorescent screen emits photons that travel down fiber optics and are registered by the CCD. The data can be thought of in a theoretical sense as a set of Poisson counts where each pixel measures the number of photons that "arrive". In Figure 1.1, the scattering angle 2θ is converted to the measured difference between the vector of incidence and the scattering vector, $\mathbf{q}$, or as $\mathbf{s}$, which is defined as $\mathbf{s} = 2\pi\mathbf{q}$. Both $\mathbf{s}$ and $\mathbf{q}$ are proportional to the radial distance observed on the sensor and as such are often treated as analogous quantities for modeling purposes. When discussing SAXS data, the term "angle" is traditionally used interchangeably with radial distance.

One of the simplifying assumptions in the SAXS methodology is the presence of radial symmetry as a result of the molecule rapidly tumbling in solution. The assumption states that molecules in solution take on all possible orientations with equal probability. This tumbling motion causes the SAXS experiment to "see" the time-averaged shell that results from rapidly rotating a three-dimensional molecule in all directions. Anecdotally, this is akin to rotating a cube on all axes quickly and observing the resulting great sphere that encompasses it. This assumption is used to justify the reduction of the data from a two dimensional sensor image to a set of one dimensional radial means and radial variances about some defined center. Figure 1.2 is an example of the sensor data with the corresponding radial mean intensity curve, shown here in log form. The sensor data appears radially symmetric with the white area representing parts of the experimental setup that block the X-rays from hitting the sensor. An optical center is determined and the data are radially averaged at fixed scattering angles to produce the radial mean intensity curve. SAXS experiments generally

consist of ensembles of concentrations and exposure times, thus producing multiple pairs of radial mean-variance curves. There does not appear to be a unified methodology to handle all the data from these ensemble experiments, with typical analyses consisting of *ad hoc* exploration of the mean curves.

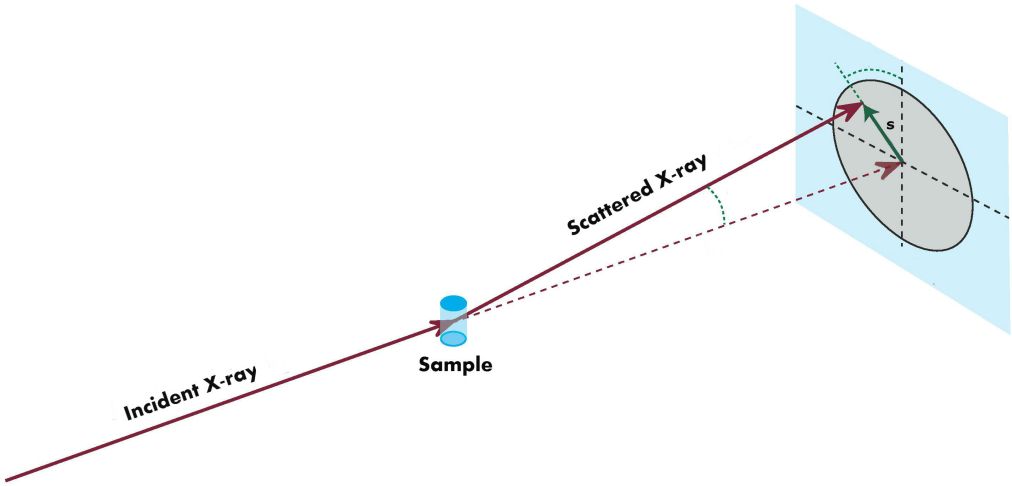


Figure 1.1: SAXS experiment consisting of an X-ray source, a sample, and a detector mechanism containing a phosphorescent screen and a CCD sensor for measuring intensity. A modified version of this figure is included in Breidt et al. (2012).

1.3 Introduction to Semiparametric Regression

The term *semiparametric regression* is commonly used to describe the class of regression models that combine traditional parametric regression with flexible nonparametric components into a cohesive framework. Parametric models, while sometimes rigid and difficult to fit, provide easy-to-understand inferential methods that can lead to significant insight to relationships in the data. Nonparametric models come in many different varieties, most of which are extremely flexible and able to fit complex relationships in a sensible manner. However, this flexibility often comes at the cost of parameter interpretability. The seminal text on semiparametric regression is Ruppert et al. (2003). The results presented here rely on their general methodology of representing semiparametric regression models as general-

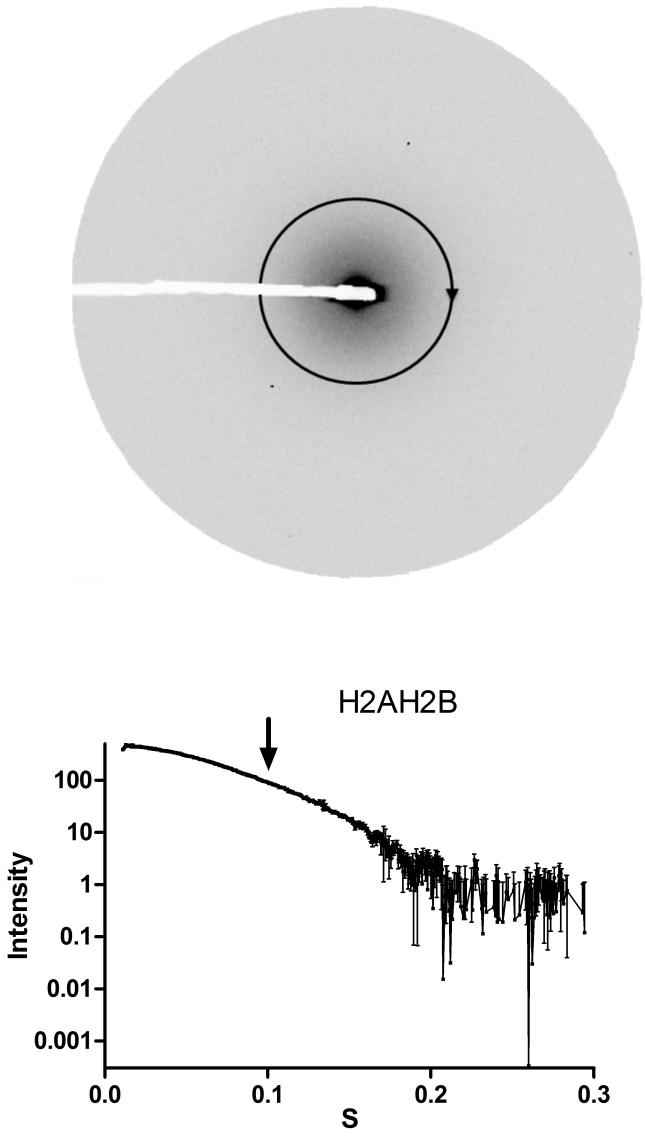


Figure 1.2: The top panel corresponds to the raw two-dimensional sensor data for a single SAXS exposure. The bottom panel is the corresponding radially averaged mean intensity data on the log scale. The arrows correspond to the radial reduction of values about a fixed annulus in the top panel and their corresponding log mean intensity in the bottom panel.

izations of linear mixed models. We present a brief introduction to this methodology here in order to provide context for our results but the reader is directed to Ruppert et al. (2003) for a more in-depth primer on the topic.

Let y_i be the response associated with the ith observation and let x_i be the ith value of some covariate of interest. A basic semiparametric regression model is

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta} + f(x_i) + \epsilon_i, \tag{1}$$

where $\mathbf{X}_i^T \boldsymbol{\beta}$ corresponds to a parametric relationship for some set of known covariates $\mathbf{X}_i^T$ with unknown regression parameters $\boldsymbol{\beta}$. The vector $\mathbf{X}_i^T$ is ith row vector of a design matrix $\mathbf{X}$. The term $f(x_i)$ represents the functional relationship between y_i and x_i that is not explained by the parametric terms. The error term ϵ_i is assumed to follow a $\mathcal{N}(0, \sigma^2)$ distribution. While there are numerous methodologies for nonparametric smoothing (wavelets, local methods, series based smoothers, neural networks, etc.), we focus on modeling $f(x)$ through a low-rank penalized spline basis expansion. This method treats $f(x)$ as a linear combination of a set of basis functions $\{B_k(x_i)\}_{k=1}^K$ over fixed, known knots $\kappa_1, \kappa_2, \ldots \kappa_K$ where $K << N$, the overall sample size. Under this expansion, the model becomes

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \sum_{k=1}^{K} b_k B_k(x_i) + \epsilon_i. \tag{2}$$

As described in Ruppert et al. (2003), there is a useful relationship between penalized spline regression and traditional mixed models. Under this framework, we treat $\boldsymbol{\beta}$ as a vector of fixed effect parameters and $\mathbf{b} = (b_1, b_2, \ldots, b_K)^T$ as a vector of random effects with common variance. This means that for each k,

$$b_k \sim \mathcal{N}\left(0, \sigma_b^2\right). \tag{3}$$

The "penalty" portion of the penalized spline regression is controlled automatically by the ratio of the model variance to the random effect variance, σ^2/σ_b^2. Large σ_b^2 corresponds to more flexibility in the nonparametric portion of the model. As $\sigma_b^2 \to 0$, the regression model approaches the fixed effect linear model.

Using this representation, we gain access to the well-documented toolset associated with mixed models for inference, model selection, hypothesis testing, etc. The availability of off-the-shelf software for both frequentist and Bayesian methodologies make these models very simple to implement. Additional parametric components are handled through the addition of new fixed effect parameters to the vector $\boldsymbol{\beta}$. New nonparametric terms can be included through an additional basis expansion controlled by an additional variance parameter. For example, one may want to build upon the model from (1) to include an additional nonparametric relationship of the form

$$y_i = \mathbf{X}_i^T\boldsymbol{\beta} + f(x_i) + g(z_i) + \epsilon_i. \tag{4}$$

Here z_i would be an additional covariate which we believe has a functional relationship with the response y_i. The term $g(z_i)$ can be expanded according to the basis $\left\{B_l^Z(z_i)\right\}$ over some knot set $\tilde{\kappa}_1, \tilde{\kappa}_2, \ldots, \tilde{\kappa}_L$. This results in the mixed model

$$\begin{aligned} y_i &= \mathbf{X}_i^T\boldsymbol{\beta} + \sum_{k=1}^{K} b_k B_k(x_i) + \sum_{l=1}^{L} c_l B_l^Z(z_i) + \epsilon_i \\ b_k &\sim \mathcal{N}\left(0, \sigma_b^2\right) \quad k = 1, \ldots, K \\ c_l &\sim \mathcal{N}\left(0, \sigma_c^2\right) \quad l = 1, \ldots, L. \end{aligned} \tag{5}$$

The general structure for these models is

$$y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \mathbf{Z}_{1i}^T\mathbf{b}_1 + \cdots + \mathbf{Z}_{Li}^T\mathbf{b}_L + \epsilon_i$$

$$\mathbf{b}_l \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{b_l}\mathcal{I}_{K_l}\right) \;\; \forall\, l = 1, \ldots, L$$
$$\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right). \tag{6}$$

Written in matrix form, this becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{b}_1 + \cdots + \mathbf{Z}_L\mathbf{b}_l + \epsilon$$
$$\mathbf{b}_l \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{b_l}\mathcal{I}_{K_l}\right) \;\; \forall\, l = 1, \ldots, L$$
$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\mathcal{I}_N\right). \tag{7}$$

1.3.1 Basis Functions

The choice of the set of basis functions $\{B_k(x)\}$ depends on the application at hand and on subjective choice. There are many notable bases available, including truncated polynomials, radial basis functions, thin-plate splines, B-splines (Eilers and Marx, 1996), and O'Sullivan splines (Wand and Ormerod, 2008). The work presented here does not focus on the choice of basis functions since the mixed model framework allows for flexibility in that regard. We choose to work primarily with truncated polynomial basis functions given their simple implementation and conceptual nature. These basis functions take the form

$$1,\; x,\; x^2, \ldots,\; x^p, (x-\kappa_1)^p_+, \ldots, (x-\kappa_K)^p_+ \tag{8}$$

where p is a positive integer representing the degree of the polynomial and

$$(x-\kappa_1)^p_+ = \begin{cases} (x-\kappa_k)^p & x > \kappa_k \\ 0 & x \leq \kappa_k. \end{cases} \tag{9}$$

Under this basis, the nonparametric model $y = f(x) + \epsilon$ is written as

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \sum_{k=1}^{K} b_k \left(x_i - \kappa_k\right)_+^p + \epsilon_i, \tag{10}$$

where $1, x, \ldots, x^p$ are fixed effects and the truncated polynomial terms are random effects. In particular, truncated quadratic splines ($p = 2$) are used throughout this work to estimate nonparametric relationships that are believed to have continuous first derivatives and to be inherently smooth.

The use of these models to describe data arising from SAXS experiments will be briefly touched upon in Section 2.1. As mentioned earlier, mixed models have convenient representations under both frequentist and Bayesian perspectives. While off-the-shelf mixed model software is sufficient for performing standard semiparametric regression of this type, a Bayesian implementation via Gibbs sampling is detailed in Section 3.1. This Bayesian framework better suits our extensions to standard semiparametric regression that are detailed throughout this paper, so it is favored for the sake of consistency.

CHAPTER 2

SEMIPARAMETRIC MODELS FOR SAXS DATA

2.1 Traditional Semiparametric Regression

Traditional analysis of experimental SAXS data revolves around the fitting of single or multiple one-dimensional curves consisting of radially averaged log intensity values. As described in Section 1.2, these data are often preprocessed through some "black box" mechanism that results in highly smoothed behavior, which can be described with a semiparametric regression model.

Semiparametric regression is particularly suited for modeling SAXS data because of the ease of including additional experimental covariates such as exposure time and concentration while modeling the underlying radial mean curves in a nonparametric fashion. Scientific knowledge sometimes suggests a known form for including these covariates (e.g. concentration is believed to have an additive effect on log intensity). Other times, parametric-by-nonparametric or fully nonparametric interactions are required, both of which can be done simply in the mixed model framework. The inferential toolbox for mixed models allows for thorough investigation of these experimental effects as well as estimating physical structural parameters. Specific examples are found in Chapter 5.

2.2 Semiparametric Regression with Heteroskedastic Errors

One of the main assumptions made in the standard semiparametric model framework is the presence of homoskedastic errors. In notational terms, this implies σ^2 is constant for all covariate values. This is a standard assumption for regression modeling but can easily be violated in applications. Heteroskedastic (non-constant) errors can have significant

consequences for model inference. It is known that for standard ordinary least squares (OLS) regression, the estimates of the model coefficients are still unbiased but suffer from improper variance estimates (Barreto and Howland, 2005, chap. 19). Heteroskedasticity can have an even greater effect in the penalized spline framework used here since the smoothing parameter is controlled by the ratio σ^2/σ_b^2 (where σ_b^2 is the random effects variance).

Heteroskedasticity is a common problem for data arising from SAXS experiments. The right panel of Figure 1.2 is the log radial mean intensity data associated with a single SAXS exposure. Figure 1.2 displays clear heteroskedasticity about a smoothly varying underlying mean function; the variance in intensity increases as s increases. The prevailing explanation for this behavior is tied to the physical process from which the data are collected. The spatial covariate $\mathbf{s}$ corresponds to the distance from the center of the two-dimensional detector at which the intensity is measured. The rarity of photon detection emitted from the aligned phosphorescent screen increases proportionally with $\mathbf{s}$, leading to increased variation further away from the center.

There are several traditional methods for handling heteroskedastic errors in linear models. Often nonlinear transformations, such as square root or log, are useful in stabilizing the variance. However, as pointed out in Carroll and Ruppert (1988), transformations are only appropriate if $\mathrm{Var}\,(Y \mid X) = h\,(E[Y \mid X])$ where Y is the response and X is the set of covariates. Weighted least squares methods can also be used to address the heteroskedastic variance structure as well (Carroll and Ruppert, 1982a). As an alternative, we consider the idea of *variance function estimation* where the log model variance, $\log(\sigma^2)$, is modeled as some smooth function of the covariates. The standard introduction to this idea is found in Carroll and Ruppert (1988) with the groundwork being laid in Carroll and Ruppert (1982b). A unified approach to variance function estimation was first presented in Davidian and Carroll (1987). Variance function estimation is still an active field of investigation. More

recent examples of work on this topic can be found in Opsomer et al. (1999) and Crainiceanu et al. (2007).

Consider the simple regression model that includes a variance function term

$$\begin{aligned} y_i &= f(x_i) + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}\left(0, \sigma_i^2\right) \\ \log(\sigma_i^2) &= g(x_i) \ \ (i = 1, 2, \ldots, N). \end{aligned} \tag{11}$$

Here the log variance function, $\log(\sigma^2)$, is modeled as a smooth function $g(x)$. An error term is omitted at this level since we do not have direct observations of $\log(\sigma^2(x))$. Adopting a penalized spline formulation as described in Ruppert et al. (2003), the model takes the form for all $i = 1, 2, \ldots, N$:

$$\begin{aligned} y_i &= [1, x_i, \ldots, x_i^p, (x_i - \kappa_1)_+^p, \ldots, (x_i - \kappa_K)_+^p] \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix} + \epsilon_i \ = \ \mathbf{C_i}^T \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix} + \epsilon_i \\ \mathbf{b} &= (b_1, b_2, \ldots, b_K)^T \ \sim \mathcal{N}_K\left(0, \sigma_b^2 \mathcal{I}_K\right) \\ \epsilon_i &\overset{\text{ind.}}{\sim} \mathcal{N}(0, \sigma_i^2) \\ \log(\sigma_i^2) &= \left[1, x_i, \ldots, x_i^r, (x_i - \kappa_1^V)_+^r, \ldots, (x_i - \kappa_{K_V}^V)_+^r\right] \begin{bmatrix} \delta \\ \mathbf{c} \end{bmatrix} \ = \ \mathbf{C_{V_i}}^T \begin{bmatrix} \delta \\ \mathbf{c} \end{bmatrix} \\ \mathbf{c} &= (c_1, c_2, \ldots, c_{K_V})^T \ \sim \mathcal{N}_{K_V}\left(0, \sigma_c^2 \mathcal{I}_{K_V}\right), \end{aligned} \tag{12}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ and $\delta = (\delta_0, \delta_1, \ldots, \delta_r)^T$ are unknown parameter vectors and $\{\kappa_k\}_{k=1}^K$ and $\left\{\kappa_k^V\right\}_{k=1}^{K_V}$ are sets of fixed, known knots. The truncated polynomial spline basis is chosen for convenience. Other common choices of basis functions include radial basis functions, B-splines, and thin plate splines, depending on the application.

Define the $N \times (p + K + 1)$ and $N \times (r + K_V + 1)$ design matrices $\mathbf{C} = \left[\mathbf{C_i}^T\right]_{i=1}^N$ and $\mathbf{C}_V = \left[\mathbf{C_{V_i}}^T\right]_{i=1}^N$. Let $\boldsymbol{\theta} = (\beta^T, \mathbf{b}^T)$ and $\boldsymbol{\theta}_V = (\delta^T, \mathbf{c}^T)$. Let $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ and $\mathbf{V} = (\log(\sigma_1^2), \ldots, \log(\sigma_N^2))^T$. It is convenient to view (12) as two mixed models with fixed effects β and δ and random effects $\mathbf{b}$ and $\mathbf{c}$. The parameters of interest for the model are $\boldsymbol{\theta}, \boldsymbol{\theta}_V, \sigma_b^2$, and σ_c^2.

Equation (12) can be considered a submodel of the class of spatially adaptive penalized spline regressions with heteroskedastic errors described in Crainiceanu et al. (2007). These models allow for both the error variance (σ^2) and the random effects variance (σ_b^2) to be smoothly varying over some set of covariates. Our model is a simplification that holds σ_b^2 as a univariate parameter to be estimated. A limitation of a model of this form is that it can have trouble fitting data in regions where dramatic increases in variation coincide with strong changes in curvature. The inclusion of a spatially adaptive random effect variance improves the ability of the model to respond to changes in the curvature of the underlying function $f(x)$. The work presented here focuses on cases where the underlying first derivative of $f(x)$ appears to vary slowly and does not coincide with sharp increases of variation, as is the case for our motivating application to SAXS data.

It is relatively simple to extend (12) to allow for multiple nonparametric terms at both the mean and variance levels. This corresponds to the inclusion of additional fixed and random effect terms with different random effect variance terms. The general heteroskedastic semiparametric model can be written as

$$
\begin{aligned}
y_i &= \mathbf{X}_i^T\beta + \mathbf{Z}_{1i}^T\mathbf{b_1} + \mathbf{Z}_{2i}^T\mathbf{b_2} + \cdots + \mathbf{Z}_{Li}^T\mathbf{b_L} + \epsilon_i = \mathbf{C_i}^T\theta + \epsilon_\mathbf{i} \\
\mathbf{b_l} &\sim \mathcal{N}_{K_l}\left(\mathbf{0}, \sigma_{b_l}^2\mathcal{I}_{K_l}\right) \quad \forall\, l = 1, 2, \ldots L \\
\epsilon_i &\sim \mathcal{N}(0, \sigma_i^2) \\
\log(\sigma_i^2) &= \mathbf{X}_{V_i}^T\delta + \mathbf{Z}_{V_{1i}}^T\mathbf{c_1} + \mathbf{Z}_{V_{2i}}^T\mathbf{c_2} + \cdots + \mathbf{Z}_{V_{Mi}}^T\mathbf{c_M} = \mathbf{C}_{V_i}^T\theta_V
\end{aligned}
$$

$$\mathbf{c_m} \sim \mathcal{N}_{K_{V_m}}\left(\mathbf{0}, \sigma^2_{c_m}\mathcal{I}_{K_{V_m}}\right) \quad \forall\, m = 1, 2, \ldots M. \tag{13}$$

Here L and M correspond to the number of unique random effect variances for the mean and variance levels respectively. The regression parameters are defined as $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \mathbf{b}_1^T, \ldots, \mathbf{b}_L^T\right)^T$ and $\boldsymbol{\theta}_V = \left(\boldsymbol{\delta}^T, \mathbf{c}_1^T, \ldots, \mathbf{c}_M^T\right)^T$. The fixed effect matrices $\mathbf{X}$ and $\mathbf{X}_V$ have dimension $N \times p+1$ and $N \times r+1$ respectively. The random effect matrices $\mathbf{Z}_l$ and $\mathbf{Z}_{V_m}$ have dimensions $N \times K_l$ and $N \times K_{V_m}$ respectively. This results in overall design matrices $\mathbf{C}$ and $\mathbf{C}_V$ with dimensions $N \times p + 1 + \sum_{l=1}^{L} K_l$ and $N \times r + 1 + \sum_{m=1}^{M} K_{V_M}$.

2.3 Semiparametric Regression for Joint Mean-Variance Data Structures

We now extend the concept of heteroskedastic semiparametric regression to handle a novel data structure unique to SAXS experiments. Recall the process of obtaining a SAXS exposure described in Section 1.2. A solution of the sample of interest is bombarded by a high powered X-ray beam which are scattered onto a two-dimensional detector. Traditional SAXS data analysis does not consider this two-dimensional response surface. Rather, this surface is reduced to a one-dimensional collection of the radially averaged intensity values. The left panel of Figure 1.2 is an example of a two-dimensional response surface from a SAXS experiment. We note that there are some important assumptions that underlie the reduction of the two dimensional data set to a one dimensional representation, which are not obviously satisfied, while at the same time, the goal of inferring structural information in three dimensions is more difficult when the data is one dimensional rather than two dimensional.

Radial dimension reduction often produces in two sets of radial summary statistics. The first is the radial sample mean intensity data. This is what is traditionally analyzed in SAXS experiments and is the focus of the models presented in Sections 2.1 and 2.2. Additionally, we may have a set of radial sample variances corresponding to these radial sample means.

In other words, for each fixed distance from a defined center, the data-collection mechanism traverses the corresponding annulus and obtains the response pair $\left(\bar{I}_i, t_i^2\right)$. Often there is a degree of artificially induced smoothness associated with these measures. This at least partially stems from the "black box" pre-processing of the data that occurs at the source. The details on these processes are rather vague. Most likely there occurs some 2D smoothing of the raw sensor data and the observed response pairs are generated using the fitted values. Since inferring structural information is an inverse problem, reducing the raw data to a one dimensional representation only makes the ill-posed nature of the problem worse.

2.3.1 Radial Variance Versus Model Variance

Before we propose a modeling paradigm, it is important to understand what exactly we are investigating. The broad goal of a unified framework that includes both radial mean and radial variance data is to gain additional insight that leads to better estimation of physical characteristics of the molecule in question. The established literature for estimating molecular structural characteristics depends on radial mean data. With the addition of the radial variance responses, we have two distinct forms of uncertainty which have impact on this data.

The first source of uncertainty information is the radial variance data themselves. For a fixed distance s_i away from the center, let $\{I_{ij}\}_{j=1}^{N_i}$ be the intensity values that fall on the corresponding annulus. These values can be thought of as the "original" values from the two-dimensional response surface. The radial sample mean and radial sample variance pair are

$$\begin{aligned} \bar{I}_i &= \frac{1}{N_i} \sum_{j=1}^{N_i} I_{ij} \\ t_i^2 &= \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(I_{ij} - \bar{I}_i\right)^2 . \end{aligned} \tag{14}$$

Here t_i^2 describes the amount of variability found across a specific annulus. This can be thought of as a form of angular variability. Large values of t_i^2 potentially suggest significant variability in between intensity information about that annulus. That is, intensity curves taken about different rays emanating from the center could be significantly different from each other. In this case using the radially averaged intensity data alone may ignore information about these differences.

The second source of uncertainty information comes from the variability of the radial mean values across annuli. In Section 2.2 we discussed the implications of observed heteroskedasticity for semiparametric regression. The non-constant variance term, σ_i^2, described the variation of the responses about some underlying smooth mean function tied to a spatial covariate. In terms of the radial mean intensity data, σ_i^2 describes the variation of $\left\{\bar{I}_i\right\}_{i=1}^N$ as a function of the distance from center s_i. The importance of the distinction between t_i^2 and σ_i^2 becomes evident in subsequent sections. For convenience, σ_i^2 is referred to as the *model variance* while t_i^2 is referred to as the *radial variance*.

If the mean-level response $y_i = \bar{I}_i$, then the relationship between radial and model variances is the same as that of a population variance and the variance of the sample mean. For simplicity, let A_i refer to the annulus defined by the radial distance s_i. Denote T_i^2 as the true population variance of all intensity measures about the A_i. Recall that by definition of the mean level model found in (13), $\text{Var}(\bar{I}_i) = \sigma_i^2$. Assuming that the intensity values about A_i, $\{I_{ij}\}_{j=1}^{N_i}$, are independent and identically distributed, then

$$\sigma_i^2 = \frac{T_i^2}{N_i}$$

$$\log(\sigma_i^2) = \log(T_i^2) - \log(N_i). \tag{15}$$

We do not have explicit knowledge of the N_i, the samples size about A_i. However, we know that the N_i is directly proportional to the distance s_i. This suggests the relationship

$$\log(\sigma_i^2) = \log(T_i^2) - \log(s_i) - \log(\alpha), \tag{16}$$

where α is some constant. Since we have estimates of T_i^2, this suggests a relationship between t_i^2 and σ_i^2 of

$$\log(t_i^2) = \log(\alpha) + \log(s_i) + \log(\sigma_i^2) + u_i \tag{17}$$

where u_i corresponds to some error term.

2.3.2 Joint Modeling Framework

For a single set of radial mean and radial variance data, consider a generic model of the form

$$\begin{aligned} y_i &= f(s_i) + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}\left(0, \sigma_i^2\right) \\ \log(\sigma_i^2) &= g(s_i) \\ \log(t_i^2) &= h(s_i, \sigma_i^2) + u_i \\ u_i &\sim \mathcal{N}\left(0, \sigma_u^2\right). \end{aligned} \tag{18}$$

For now let $y_i = \bar{I}_i$. In Chapter 5, we discuss the implications of modeling $\bar{I}_i$ vs. $\log(\bar{I}_i)$. For the models described in Sections 2.1 and 2.2, we can choose either response based on personal preference. However, for the model described here, we limit ourselves to the untransformed radial mean intensity values in order to preserve the structural relationship between σ^2 and t^2. Here $f(s)$, $g(s)$, and $h(s, \sigma^2)$ represent smooth functional relationships. If the last two

lines of (18) are omitted, then this is a heteroskedastic semiparametric regression model as described in Section 2.2. The functional relationships $f(s)$ and $g(s)$ are represented according to some basis expansion that will yield fixed and random effects as before.

By including the last two lines of (18), we set forth to jointly model both radial measurements with two explicit aims. First, we treat $\log(t_i^2)$ as an additional response of interest in hopes of understanding the effect of experimental conditions on the radial variance processes associated with the data collection mechanism. Scientific understanding of this process is sparse and initial investigations have revealed the existence of unexpected experimental effects that warrant investigation. As with mean intensity measures, experimental factors such as concentration, exposure time, and sensor calibration can have dramatic effects on data quality. Detecting these effects is a primary goal of all the methodologies set forth in this work. Secondly, by treating $\log(t_i^2)$ as a function of not only radial distance but also model variance, we hope to better inform the uncertainty estimates associated with the radial mean portion of the model.

The functional relationship $h(s, \sigma^2)$ in (18) is of particular interest here. A simple $h(s, \sigma^2)$ motivated by (18) is the fixed effect model

$$\begin{aligned} \log(t_i^2) &= \eta_0 + \eta_1 \log(s_i) + \eta_2 \log(\sigma_i^2) + u_i \\ u_i &\sim \mathcal{N}\left(0, \sigma_u^2\right). \end{aligned} \tag{19}$$

More flexible nonparametric forms could be implemented for $h(s, \sigma^2)$ using the standard expansions described here. These could include additive semiparametric models ($h(s, \sigma^2) = h_1(s) + h_2(\sigma^2)$) or a two-dimensional nonparametric fit using an appropriate basis expansion (e.g., thin plate splines).

We detail a Bayesian approach for parameter estimation for (18) in Section 3.3. The method used is an extension of the hybrid Gibbs sampler with delayed rejection adaptive Metropolis (DRAM) steps used for the heteroskedastic semiparametric regression from Section 3.2. Also, we discuss a variational approximation for the joint mean-variance model when the radial variance level consists of the fixed effect structure from (19) in Section 4.4.

CHAPTER 3

BAYESIAN INFERENCE VIA MARKOV CHAIN MONTE CARLO FOR SEMIPARAMETRIC MODELS OF SAXS DATA

3.1 Semiparametric Regression via Gibbs Sampling

We begin this section by detailing a well-known Bayesian framework for mixed models that is useful in semiparametric regression problems. Recall the general mixed model from (7), rewritten here for convenience:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{b}_1 + \cdots + \mathbf{Z}_L\mathbf{b}_l + \epsilon \\ \mathbf{b}_l &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{b_l}\mathcal{I}_{K_l}\right) \;\; \forall\, l = 1, \ldots, L \\ \epsilon &\sim \mathcal{N}\left(0, \sigma^2\mathcal{I}_N\right). \end{aligned} \tag{20}$$

The parameters of interest for this model can be classified as regression parameters, $\boldsymbol{\beta}$, $\mathbf{b}_1$, $\ldots$, $\mathbf{b}_L$, and variance parameters, $\sigma^2_{b_1}, \ldots, \sigma^2_{b_L}, \sigma^2$. Prior distribution specification is required for $\boldsymbol{\beta}$ as well as all variance parameters. Well-known conjugate prior structures can be used to allow for the explicit computation of posterior parameter conditional distributions. Consider the Normal and Inverse Gamma conjugate priors

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_\beta\mathcal{I}_{p+1}\right) \\ \sigma^2_{b_l} &\sim \mathcal{IG}\left(A_{b_l}, B_{b_l}\right) \;\; \forall\, l = 1, \ldots, L \\ \sigma^2 &\sim \mathcal{IG}\left(A, B\right), \end{aligned} \tag{21}$$

where the $\mathcal{IG}(A, B)$ denotes an Inverse Gamma distribution with density function

$$f(x) = \frac{B^A}{\Gamma(A)} x^{-A-1} \exp\left(\frac{-B}{x}\right). \tag{22}$$

Define $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \mathbf{b}_1^T, \ldots, \mathbf{b}_L^T\right)^T$ as the vector of all regression parameters. Under this prior structure, the posterior conditional parameter distributions for $\boldsymbol{\theta}$, $\sigma^2_{b_1}$, ..., $\sigma^2_{b_L}$, and σ^2 are

$$\begin{aligned}
\boldsymbol{\theta} \mid \cdot &\sim \mathcal{N}\left(\frac{1}{\sigma^2}\mathbf{M}\mathbf{C}^T\mathbf{Y}, \mathbf{M}\right) \\
\mathbf{M} &= \left(1/\sigma^2\mathbf{C}^T\mathbf{C} + \Sigma_\theta^{-1}\right)^{-1} \\
\boldsymbol{\Sigma}_\theta &= \text{blockdiag}\left(\sigma^2_\beta \mathcal{I}_{p+1}, \sigma^2_{b_1}\mathcal{I}_{K_1}, \ldots, \sigma^2_{b_L}\mathcal{I}_{K_L}\right) \\
\sigma^2_l \mid \cdot &\sim \mathcal{IG}\left(A_l + \frac{K_l}{2}, B_l + \frac{\|\mathbf{b}_l\|^2}{2}\right) \quad \forall\, l = 1, \ldots, L \\
\sigma^2 \mid \cdot &\sim \mathcal{IG}\left(A_\epsilon + \frac{N}{2}, B_\epsilon + \frac{\|\mathbf{Y} - \mathbf{C}\boldsymbol{\theta}\|^2}{2}\right).
\end{aligned} \tag{23}$$

The notation $A \mid \cdot$ denotes the conditional distribution of A on all other parameters and data in the model.

Since we have known forms for the posterior parameter conditional distributions for all parameters, a MCMC algorithm based on Gibbs sampling is appropriate for posterior sampling (Casella and George, 1992). After initialization, the posterior conditional distributions are sequentially sampled using the most recent draws for all other parameters. This process is repeated for some fixed number of iterations. As discussed in Casella and George (1992), repeatedly sampling from alternating posterior conditional distributions for a sufficiently large number of iterations will yield an appropriate empirical representation of the posterior parameter distribution. Parameter estimates are constructed from the parameter chains representing the empirical posterior conditional distributions.

3.1.1 Single Curve Example

To illustrate semiparametric regression via the Gibbs sampler described above, we consider data arising from the underlying smooth function

$$m(x) = 50\phi\left(\frac{x - 3.5}{1.25^2}\right) + 50\phi\left(\frac{x - 6.5}{1.25^2}\right), \tag{24}$$

where $\phi(z)$ is the density function of the standard Normal distribution, $\mathcal{N}(0, 1)$. We simulated $N = 200$ responses over the range of x according to the relationship $y = m(x) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$.

As described in (10), we use a truncated quadratic penalized spline basis so that the model takes the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sum_{k=1}^{K} b_k \left(x_i - \kappa_k\right)_+^2 \tag{25}$$

over a set of $K = 10$ knots at equally spaced quantiles of $\{x_i\}$, ranging from from 1% to 99%. The Gibbs sampler described above is run for $10,000$ iterations with a burn-in of 1000. The resulting smooth fit, $\hat{m}(x)$ is shown along with 95% credible bounds in Figure 3.1.

This MCMC method handles the semiparametric regression model quite well. Analysis of the parameter trace plots in Figure 3.2 indicates appropriate mixing of the posterior parameter samples and does not raise any red flags regarding these results.

3.1.2 Multiple Curve Example

The flexibility of this regression framework allows for straightforward handling of parametric-by-nonparametric and fully nonparametric interactions among covariates. Additional model complexity, through the addition of the necessary fixed and random effects, has minimal

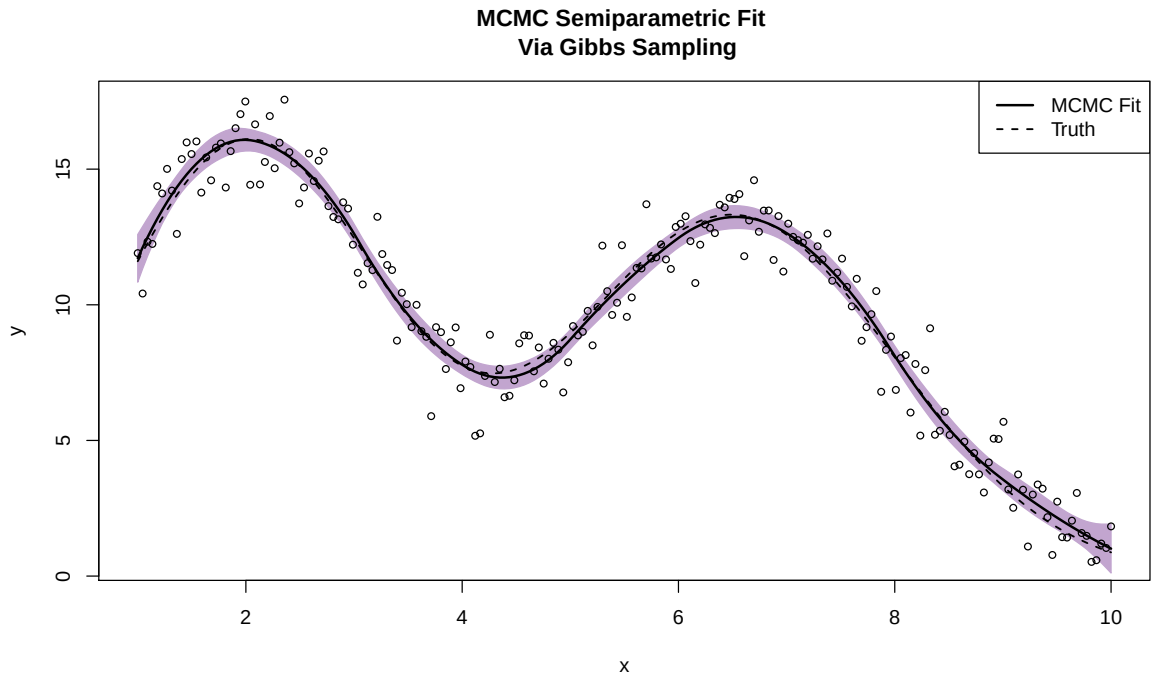


Figure 3.1: Truncated quadratic spline model for simulated data. The data are simulated with true mean function $m(x) = 50\phi\left(\frac{x-3.5}{1.25^2}\right) + 50\phi\left(\frac{x-6.5}{1.25^2}\right)$ and standard Gaussian error. Shaded regions correspond to 95% pointwise credible bounds generated from the estimated fits of $\hat{\mathbf{y}}$. The Gibbs sampler was run for $10,000$ iterations with a 1000 iteration burn-in.

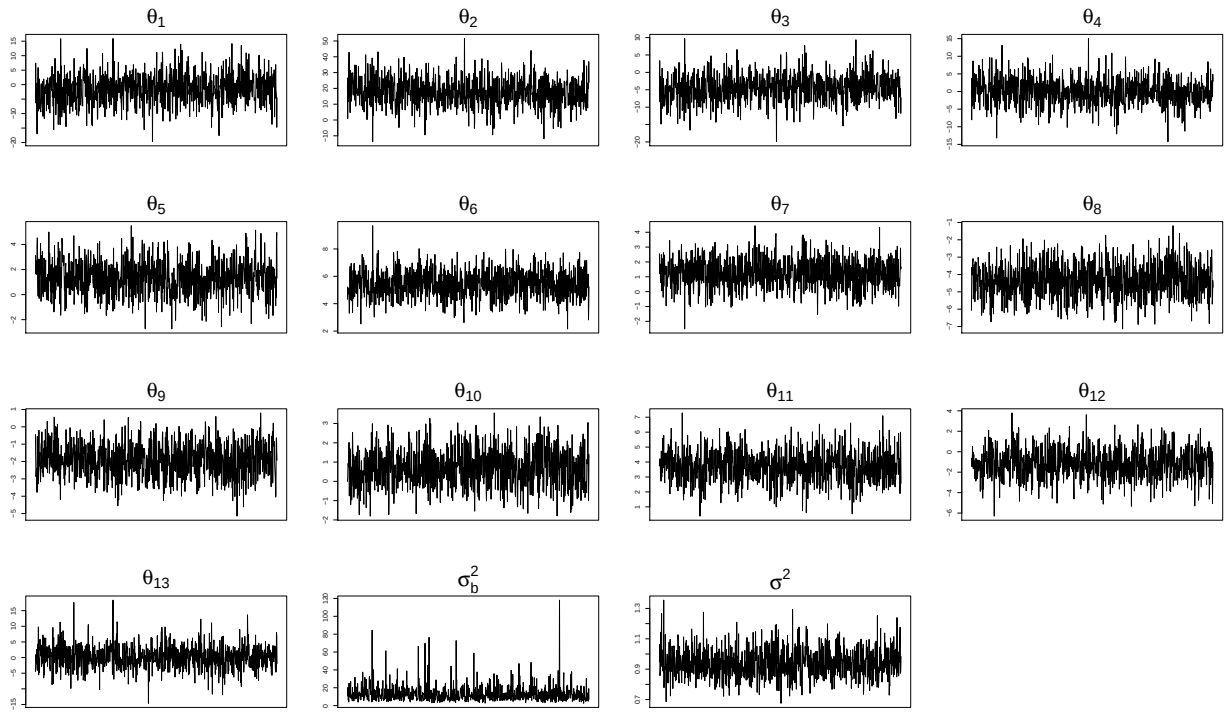


Figure 3.2: Trace plots of parameter chains of the Gibbs sampler for all elements of $\boldsymbol{\theta}$, σ_b^2, and σ^2. The plots depict the 9000 iterations post burn-in.

effect on the computational mechanism used to estimate these parameters. We now provide a multiple curve example that demonstrates this property.

Consider two smooth mean functions over the region $[0, 10]$

$$\begin{aligned} m_1(x) &= 50\phi\left(\frac{x-3.5}{1.25^2}\right) + 50\phi\left(\frac{x-6.5}{1.25^2}\right) \\ m_2(x) &= 50\phi\left(\frac{x-3.5}{1.25^2}\right) + 50\phi\left(\frac{x-(6.5+S))}{2^2}\right). \end{aligned} \tag{26}$$

As before, $\phi(z)$ is the standard Normal density function. Here the variable S corresponds to some shift mechanism in the second Gaussian density function. For the data in Figure 3.3, we generate 200 responses from equally spaced values of x across the region for each mean function. Standard Normal error is added to each mean function to simulate our responses.

An appropriate model for this data would include a parametric-by-nonparametric interaction term to describe the relationship between the two mean functions. Define S_i as being 0 for data generated from $m_1(x)$ and $S_i = 1.5$ otherwise. Let $\kappa_1, \ldots, \kappa_{10}$ be the same knot set used in Section 3.1.1. We fit the data using the model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 S_i + \beta_4 x_i S_i + \beta_5 x_i^2 S_i i \\ &\quad + \sum_{k=1}^{K} b_{1k} \left(x_i - \kappa_k\right)_+^2 + \sum_{k=1}^{K} S_i b_{2k} \left(x_i - \kappa_k\right)_+^2 + \epsilon_i \\ \mathbf{b}_{1k} &\sim \mathcal{N}\left(0, \sigma_{b_1}^2\right) \qquad \forall\, k = 1, \ldots K \\ \mathbf{b}_{2k} &\sim \mathcal{N}\left(0, \sigma_{b_2}^2\right) \qquad \forall\, k = 1, \ldots K \\ \epsilon_i &\sim \mathcal{N}\left(0, \sigma^2\right). \end{aligned} \tag{27}$$

Figure 3.3 shows how well the model fits the data in the presence of this nonlinear interaction effect. The Gibbs sampler worked quite well in generating acceptable parameter estimates. Additional diagnostics and trace plots are omitted here for the sake of brevity since, as stated

before, the behavior of this computational procedure is well understood for standard models such as the mixed model used here.

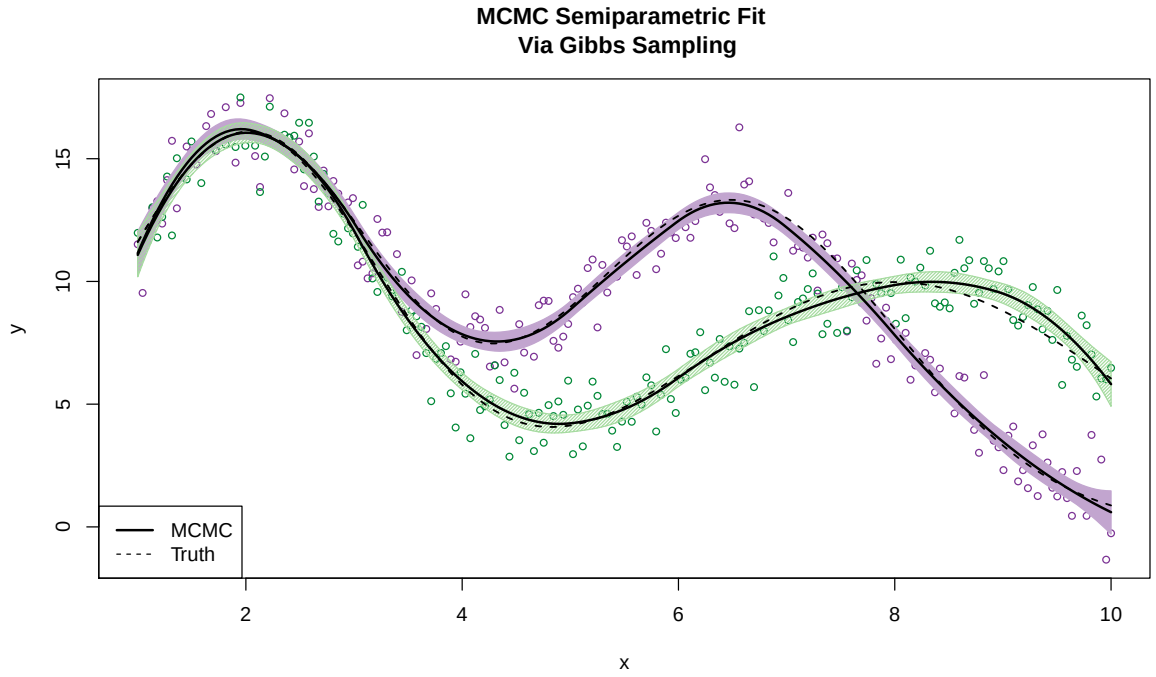


Figure 3.3: MCMC fit of semiparametric interaction model. The data are simulated with true mean functions $m_1(x)$ or $m_2(x)$ in (26), both with standard Gaussian error. Shaded regions correspond to 95% credible bounds. The Gibbs sampler was run for $10,000$ iterations with a 1000 iteration burn-in.

3.2 Heteroskedastic Semiparametric Regression via Hybrid Gibbs Sampling

The hierarchical nature of the heteroskedastic semiparametric regression model described in (13) lends itself naturally to a Bayesian approach. As with the model described in Crainiceanu et al. (2007), we use the conjugate prior structure

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\beta^2 \mathcal{I}_{p+1}\right)$$
$$\boldsymbol{\delta} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\delta^2 \mathcal{I}_{r+1}\right)$$

$$\sigma^2_{b_l} \sim \mathcal{IG}\left(A_{b_l}, B_{b_l}\right) \;\; \forall\, l = 1, \ldots, L$$
$$\sigma^2_{c_m} \sim \mathcal{IG}\left(A_{c_m}, B_{c_m}\right) \;\; \forall\, m = 1, \ldots, M, \tag{28}$$

where fixed effect coefficient parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are modeled with multivariate Gaussians and random effect variance terms $\sigma^2_{b_l}$ and $\sigma^2_{c_m}$ are modeled with Inverse Gamma distributions.

Crainiceanu et al. (2007) provides an excellent discussion on the choice of hyperparameters for both the fixed effect parameters as well as for the variance terms. It is standard practice to use a Gaussian prior structure for the fixed effect parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ as well as assuming that the components of each vector are independent. Traditionally σ^2_β and σ^2_δ are chosen to correspond to diffuse Gaussian distributions, usually on the order of 10^6. The rate and scale hyperparameters for the Inverse Gamma priors are also chosen to correspond to non-informative priors, usually with values on the order 10^{-4}. Crainiceanu et al. (2007) discusses this choice along with the challenges that can arise as a consequence of the numerical scale of the underlying spatial covariate x and the reader is referred to that paper for more in-depth analyses of hyperparameter choice. All models presented in this work had sensitivity analyses performed to ensure that our estimated posterior distributions were not being improperly informed by *a priori* assumptions.

Using the prior structure from (28), the model described in (13) yields the posterior parameter conditionals

$$\begin{aligned}
\theta \mid \cdot &\sim \mathcal{N}\left(\mathbf{M}\mathbf{C}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y}, \mathbf{M}\right) \\
\mathbf{M} &= \left(\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C} + {\boldsymbol{\Sigma}_\theta}^{-1}\right)^{-1} \\
\sigma^2_{b_l} \mid \cdot &\sim \mathcal{IG}\left(A_b + \frac{K}{2}, B_b + \frac{\|\mathbf{b}\|^2}{2}\right) \;\; \forall\, l = 1, \ldots, L \\
\sigma^2_{c_m} \mid \cdot &\sim \mathcal{IG}\left(A_c + \frac{K_V}{2}, B_b + \frac{\|\mathbf{c}\|^2}{2}\right) \;\; \forall\, m = 1, \ldots, M
\end{aligned}$$

$$p(\theta_{\mathbf{V}} \mid \cdot) \propto \exp\Big[-\frac{1}{2}\Big\{ \sum_{i=1}^{N} {\mathbf{C}_{\mathbf{V_i}}}^T \theta_{\mathbf{V}} + \sum_{i=1}^{N} \big(Y_i - {\mathbf{C}_{\mathbf{i}}}^T \theta\big)^2 \exp(-{\mathbf{C}_{\mathbf{V_i}}}^T \theta_{\mathbf{V}}) + {\theta_{\mathbf{V}}}^T {\mathbf{\Sigma}_{\theta_{\mathbf{V}}}}^{-1} \theta_{\mathbf{V}} \Big\}\Big]. \tag{29}$$

The challenge of fitting this model is due to the non-standard distribution of $\boldsymbol{\theta}_V \mid \cdot$. The distribution $p(\boldsymbol{\theta}_V \mid \cdot)$, while in the exponential family, does not represent a known distribution. A common way to handle Bayesian models that lack known posterior conditional forms for a small subset of the parameters is to use some standard MCMC method (Metropolis-Hastings, rejection sampling, etc.) to generate samples, while Gibbs sampling is performed for the remaining parameters. However, the dimensionality of $\boldsymbol{\theta}_V$ makes this problematic. The parameter $\boldsymbol{\theta}_V$ is a $r + 1 + \sum_{m=1}^{M} K_{V_m}$-dimensional vector. Even for simple models this value is often too large for simple methods to adequately explore the posterior parameter conditional distribution $\boldsymbol{\theta}_V \mid \cdot$. While there are potentially many ways to address this issue, we detail two approaches here. One approach, as done in Baladandayuthapani et al. (2005) and Crainiceanu et al. (2007), is to introduce a latent error term to the variance model that allows for the sampling from N univariate posterior conditionals, which can be accomplished through traditional Metropolis-Hastings schemes. Alternatively, one could use a more advanced sampling method for drawing from $\boldsymbol{\theta}_V \mid \cdot$. We propose the use of a delayed rejection adaptive Metropolis (DRAM) algorithm to address the multivariate sampling problem (Haario et al., 2006). We focus on the DRAM approach and present it here due to considerations for the joint mean-variance models discussed in Section 2.3 and the variational approximations described in Section 4.3. In addition to these two approaches, personal correspondence with Dr. Daniel Hernandez-Stumpfhauser has suggested the use of a multivariate "slice sampling" approach to dealing with the lack of conjugate structure in $\boldsymbol{\theta}_V \mid \cdot$ (Damlen et al., 1999). We have not investigated this idea but this approach could prove interesting for future study.

3.2.1 Delayed Rejection Adaptive Metropolis

Delayed rejection adaptive Metropolis (DRAM), as the name implies, refers to a class of sampling methodologies combining two complementary techniques: delayed rejection and adaptive Metropolis sampling. A brief introduction is provided here to provide appropriate context to our work. The reader is directed to Haario et al. (2006) for the definitive narrative on the subject.

Delayed rejection methods are based upon the idea of allowing multiple proposal distributions to be used at each iteration of a Markov chain. In traditional Metropolis-Hastings procedures, a candidate value y_1 is generated from some proposal $q_1(x_t, \cdot)$ where $x_t = X_t$, the current state of the Markov chain. This candidate is accepted with probability

$$\alpha_1(x_t, y_1) = \max\left(1, \frac{\pi(y_1)q_1(y_1, x_t)}{\pi(x_t)q_1(x_t, y_1)}\right). \tag{30}$$

Here $\pi(\cdot)$ denotes the distribution of interest for our sample (the posterior parameter distribution of $\boldsymbol{\theta}_V$ in our case). The next value in the Markov chain is defined as

$$X_{t+1} = \begin{cases} y_1 & \text{with probability } \alpha_1(x_t, y_1) \\ x_t & \text{with probability } 1 - \alpha_1(x_t, y_1). \end{cases} \tag{31}$$

A delayed rejection approach embeds an additional proposal step that is used after the rejection of y_1 and "delays" the chain from moving to the next iteration. Let y_2 be a candidate generated from the proposal distribution $q_2(x_t, y_1, \cdot)$. This candidate is accepted with probability

$$\alpha_2(x_t, y_1, y_2) = \max\left(1, \frac{\pi(y_2)q_1(y_2, y_1)q_2(y_2, y_1, x_t)[1 - \alpha_1(y_2, y_1)]}{\pi(x_t)q_1(x_t, y_1)q_2(x_t, y_1, y_2)[1 - \alpha_1(x_t, y_1)]}\right). \tag{32}$$

The next value in the Markov chain is now defined as

$$X_{t+1} = \begin{cases} y_1 & \text{with probability } \alpha_1(x_t, y_1) \\ y_2 & \text{with probability } \alpha_2(x_t, y_1, y_2) \\ x_t & \text{with probability } 1 - \alpha_1(x_t, y_1) - \alpha_2(x_t, y_1, y_2). \end{cases} \tag{33}$$

The idea behind delayed rejection is to increase the number of accepted candidates by allowing for multiple proposal distributions to be used. At each level, a different proposal can be used that can be informed by not only the current state of the chain but also the rejected candidate values from all subsequent levels. While a two-level proposal scheme is presented here, it is generalizable to an arbitrary number of proposal distributions. A common choice is to have Gaussian proposals with covariance scaling as one moves through the scheme.

Adaptive MCMC methods traditionally aim to leverage on-line information about the parameter chain in order to tune the proposal distribution to increase chain efficiency (Haario et al., 2001, 2006). The form implemented in our work tunes the covariance matrix of a multivariate Gaussian proposal distribution at fixed interval length. After some initialization, the covariance matrix updates take the form

$$\boldsymbol{\Sigma}_n = s_d \text{Cov}\left(X_1, X_2, \ldots, X_{n-1}\right) + s_d e \mathcal{I}_D, \tag{34}$$

where s_d is some scaling factor and e is a small user-defined constant. Adaptation occurs at predefined intervals. Covariance updating of this form violates the Markovian properties of the chain but it can be shown that chains following this scheme are ergodic (Haario et al., 2001, 2006).

DRAM computational schemes have been shown to be particularly effective at dealing with multivariate sampling problems, often in cases where the target distribution is far from a known form. In the context of heteroskedastic semiparametric regression, a two-stage DRAM step will be used to draw samples from $\boldsymbol{\theta}_V \mid \cdot$ while samples are drawn from the other posterior parameter conditional distributions using their known forms.

3.2.2 Single Curve Examples

To illustrate this methodology, we now consider simulated data with a known mean and variance function. Figure 3.4 shows $N = 200$ data points simulated from a true mean function of $m(x) = -\frac{1}{8}(x-5)^3 + x$ under the variance functions $v_1(x) = \left(\frac{1}{4}x + \frac{1}{2}\right)^3$ and $v_2(x) = \exp\left\{\frac{(x-5)^2}{5}\right\}$ over the region $x \in [0, 10]$.

A truncated quadratic spline model of the form found in (12) ($p = 2$), written as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+^2 + \epsilon_i \\ \log(\sigma_i^2) &= \delta_0 + \delta_1 x_i + \delta_2 x_i^2 + \sum_{k=1}^{K_V} c_k (x_i - \kappa_k^V)_+^2, \end{aligned} \tag{35}$$

is used to for both data sets. Here we use 10 equally spaced knots over $[0, 10]$ for both the mean and variance levels (K =10,000$K_V = 10$).

The MCMC procedure described above was run for $10,000$ iterations with a burn-in of 1000 for both data sets. For the DRAM procedure, we used a dual-stage Gaussian proposal scheme with covariance scaling by a factor of 100 upon first-level rejection. The adaptive interval length used for updating the proposal covariance is 100 iterations. The defined scaling factor s_d is set to $2.4/\sqrt{N_{par}}$, where N_{par} is the dimension of the parameter $\boldsymbol{\theta}_V$ (Haario et al., 2006).

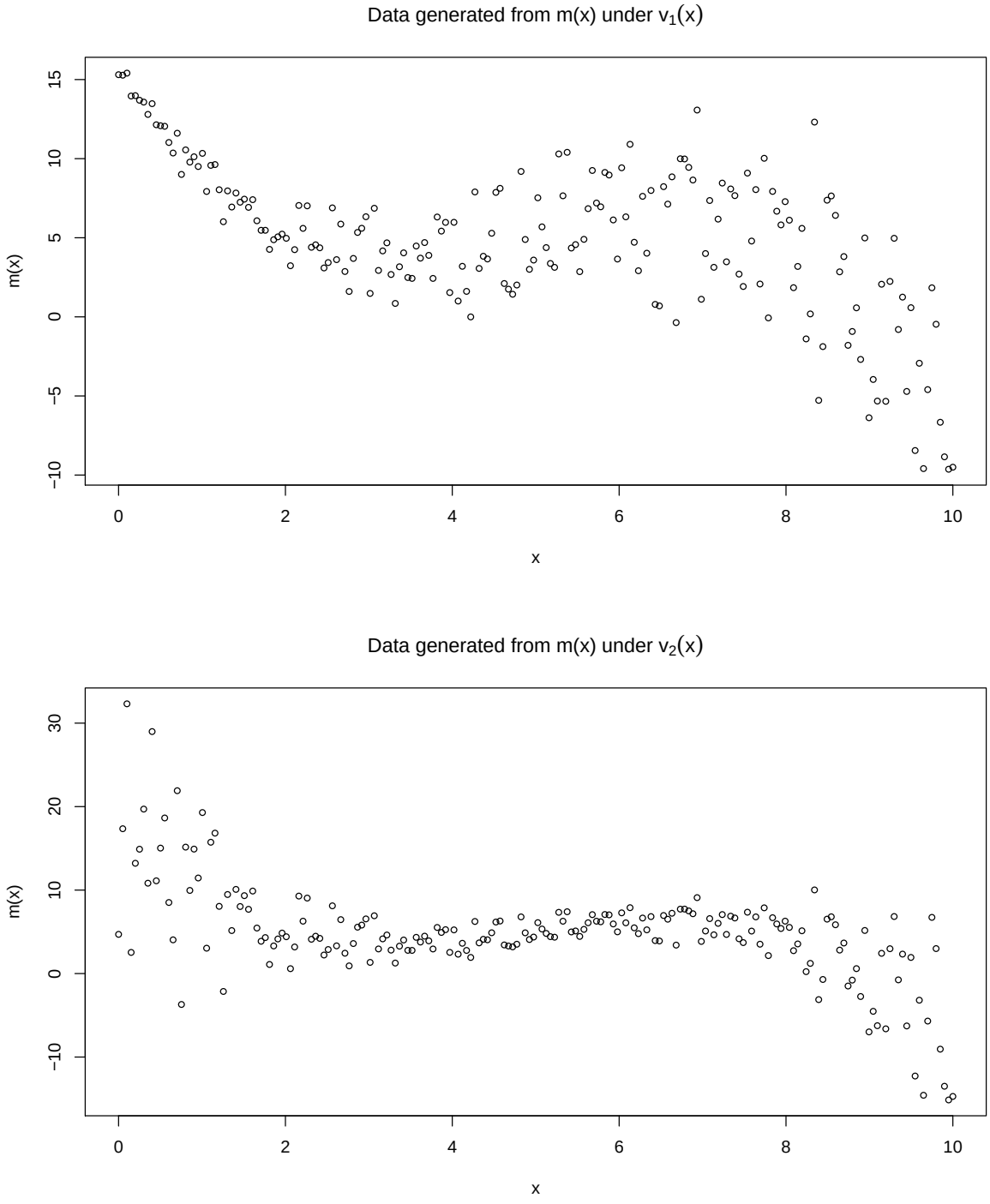
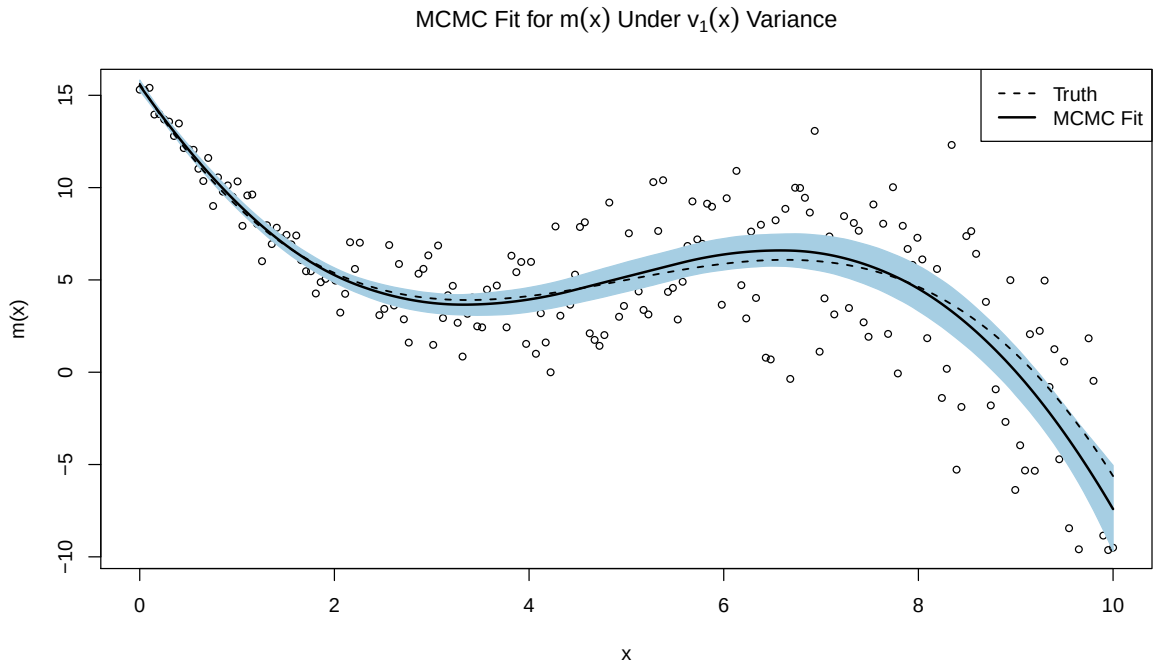


Figure 3.4: Simulated heteroskedastic data with mean function $m(x) = -\frac{1}{8}(x-5)^3 + x$. The top frame corresponds to a true variance function $v_1(x) = \left(\frac{1}{4}x + \frac{1}{2}\right)^3$ while the bottom frame corresponds to $v_2(x) = \exp\left\{\frac{(x-5)^2}{5}\right\}$.

Figure 3.5 shows the resulting fits of the simulated observations from $m(x)$ under $v_1(x)$ and $v_2(x)$ respectively. The estimated mean functions matches up closely with $m(x)$. The shaded regions represent 95% credible bounds. These regions better reflect the observed variance structure of the data than a homoskedastic model.

Apart from the simulated mean data, it may be of interest to study the estimated log variance function that drives the non-constant variance structure. Note that these variance functions are implicitly latent and never directly observed. All inference about them comes from the residual structure of the mean level model. Figure 3.6 contains the estimates of $\log(v_1(x))$ and $\log(v_2(x))$. The fits detailed in this figure raise some interesting questions. While the estimated log variance functions match up with the general structure of $\log(v_1(x))$ and $\log(v_2(x))$, there are some coverage issues with the 95% credible bounds that are notable. Most likely, the lack of fit of the log variance functions is tied to lack of direct observations of $\log(v(x))$ for both simulations. Figures 3.7 and 3.8 show the mean and log variance fits $N = 800$ simulated data points as opposed to $N = 200$. As expected, the increase in sample size increases the quality of the estimated mean function. This increase in data seems to improve the estimated log variance function but there are still some minor lack-of-fit problems that persist.

Alternatively, the difficulties in fitting $\log(v(x))$ could be tied to the computational scheme used to estimate $\boldsymbol{\theta}_V$. Even a simple two-stage DRAM method requires manual tuning of adaptive interval length, start values, covariance scaling factors, etc. Appropriate diagnostic measures are needed for assessing the appropriateness of our estimate of $\boldsymbol{\theta}_V$. For the MCMC procedure for $N = 200$, we observed acceptance rates after burn-in between 11% and 15%. Figure 3.9 is the trace plot for $\boldsymbol{\theta}_V$ associated with estimating $\log(v_1(x))$. While the overall acceptance rate is adequate, Figure 3.9 suggests that the marginal chains have not converged. However, the sluggishness in the marginal chains of $\boldsymbol{\theta}_V$ may be partially due to the fact that



$10,000$

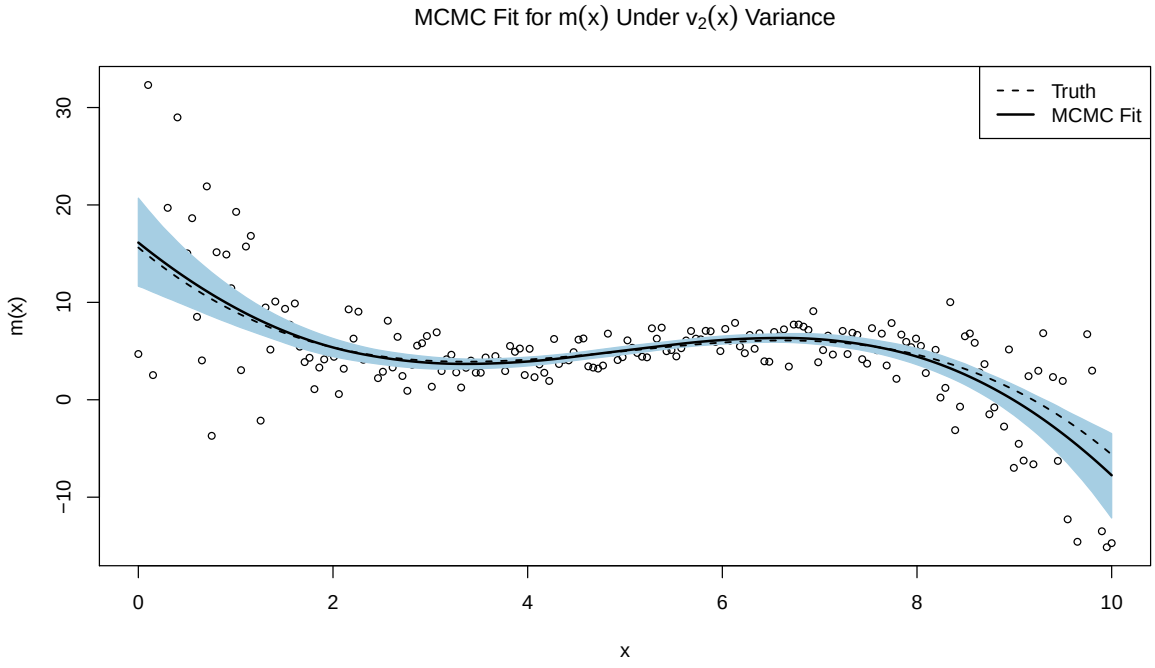


Figure 3.5: MCMC estimated mean curves from (35). The MCMC procedure was run for $10,000$ iterations iterations with a burn-in of 1000.

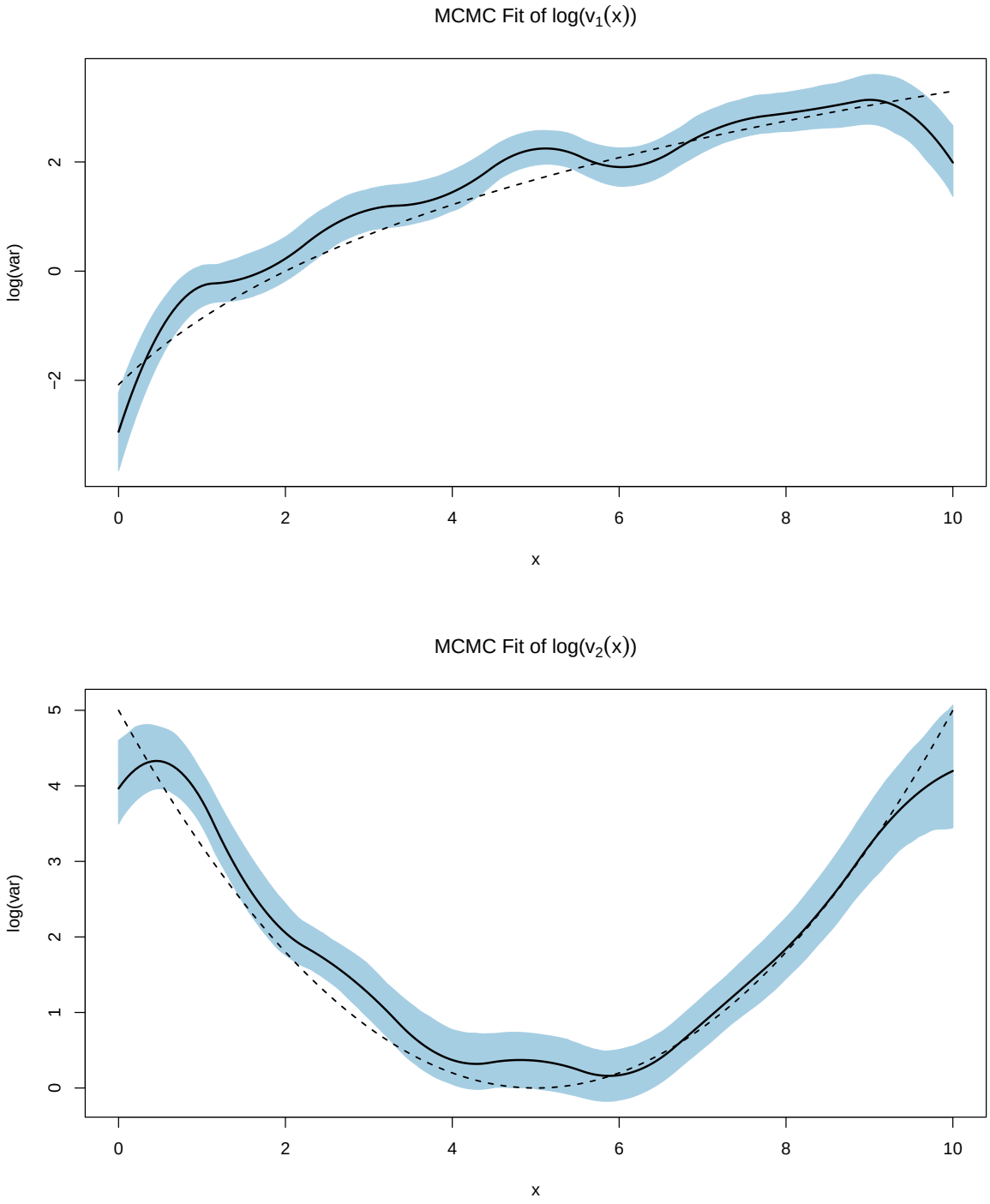


Figure 3.6: MCMC estimated log variance curves from (35). The MCMC procedure was run for $10,000$ iterations with a burn-in of 1000.

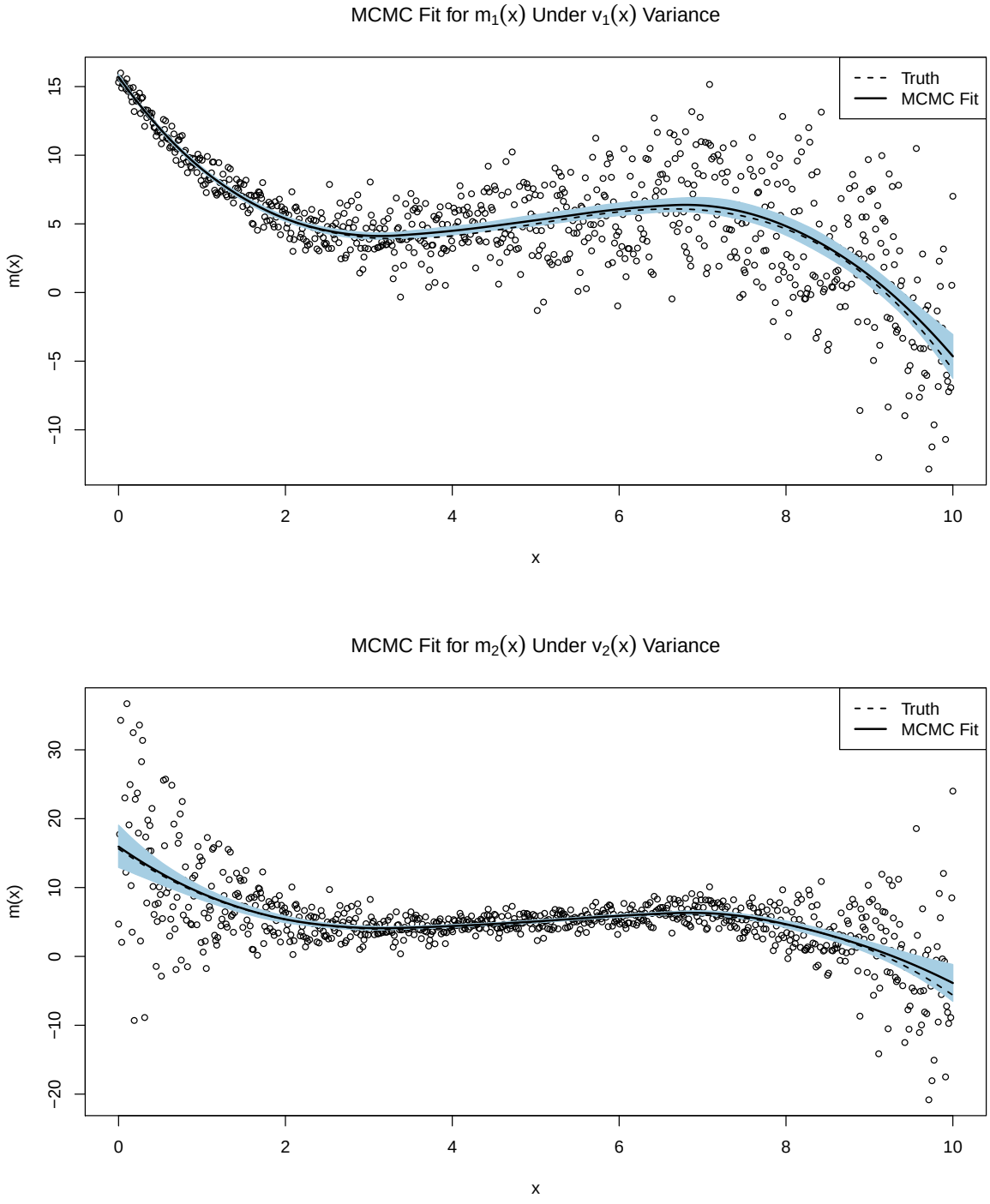


Figure 3.7: MCMC estimates of mean functions for (35) for large sample case ($N = 800$).

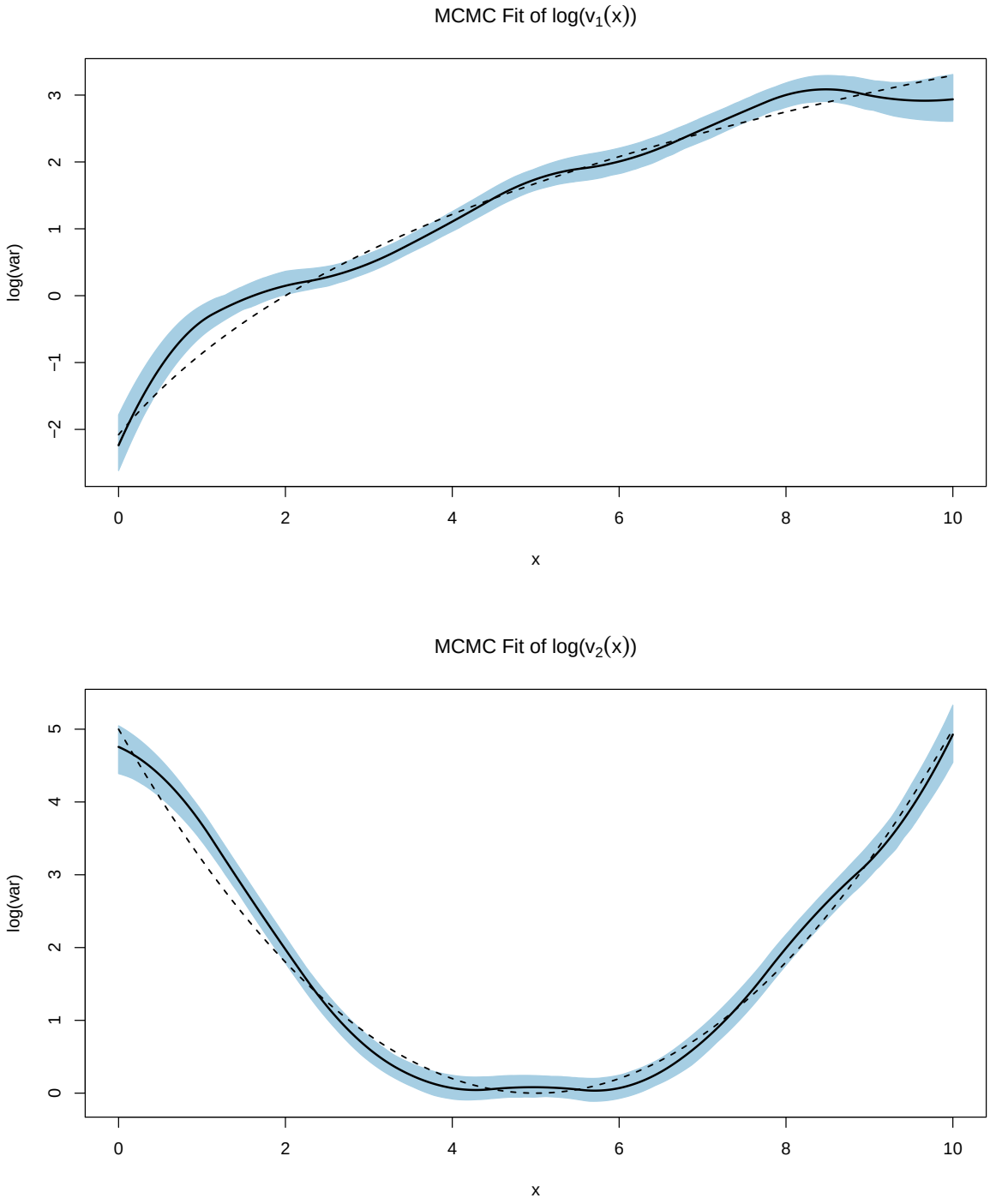


Figure 3.8: MCMC estimates of (35) log variance functions for large sample case ($N = 800$).

they are a set of basis coefficients. Analyzing them marginally may not be appropriate for assessing true convergence of the method. To compare, Figure 3.10 contains pointwise trace plots for $\hat{v}(x)$, the estimated log variance curve defined as $\mathbf{C}_V\boldsymbol{\theta}_V$. These chains seem to demonstrate better convergence properties than the marginal chains of $\boldsymbol{\theta}_V$ alone. This is not surprising since it is the variance terms $\sigma_i^2 = \exp\left(\mathbf{C}_{V_i}^T\boldsymbol{\theta}_V\right)$ that actually have representations in the mean level model.

The same procedure was repeated for 50000 iterations with a burn-in of 5000 to test the influence of a longer run-time. The resulting log variance estimates are displayed in Figure 3.11. Dramatic increases of the run length appear to improve the coverage properties of the $v(x)$ estimates. The estimates of the mean functions are virtually identical to the 10000 iteration example and omitted for brevity. However, there is still evidence that the chain for $\boldsymbol{\theta}_V$ has not converged. At this time, we see three potential remedies to this convergence issue. First, modification of the DRAM tuning parameters and/or proposal hierarchy could potentially lead to better traversal of the posterior parameter space. Secondly, replacing the truncated polynomial basis functions with an alternative, orthogonal basis (e.g., B-splines) could allow for the marginal elements of $\boldsymbol{\theta}_V$ to move more freely, speeding up computation. Finally, extending the model to include a parameter expansion representation of the regression coefficients may also help Gelman et al. (2013). Resolving this convergence issue will be a priority for future work.

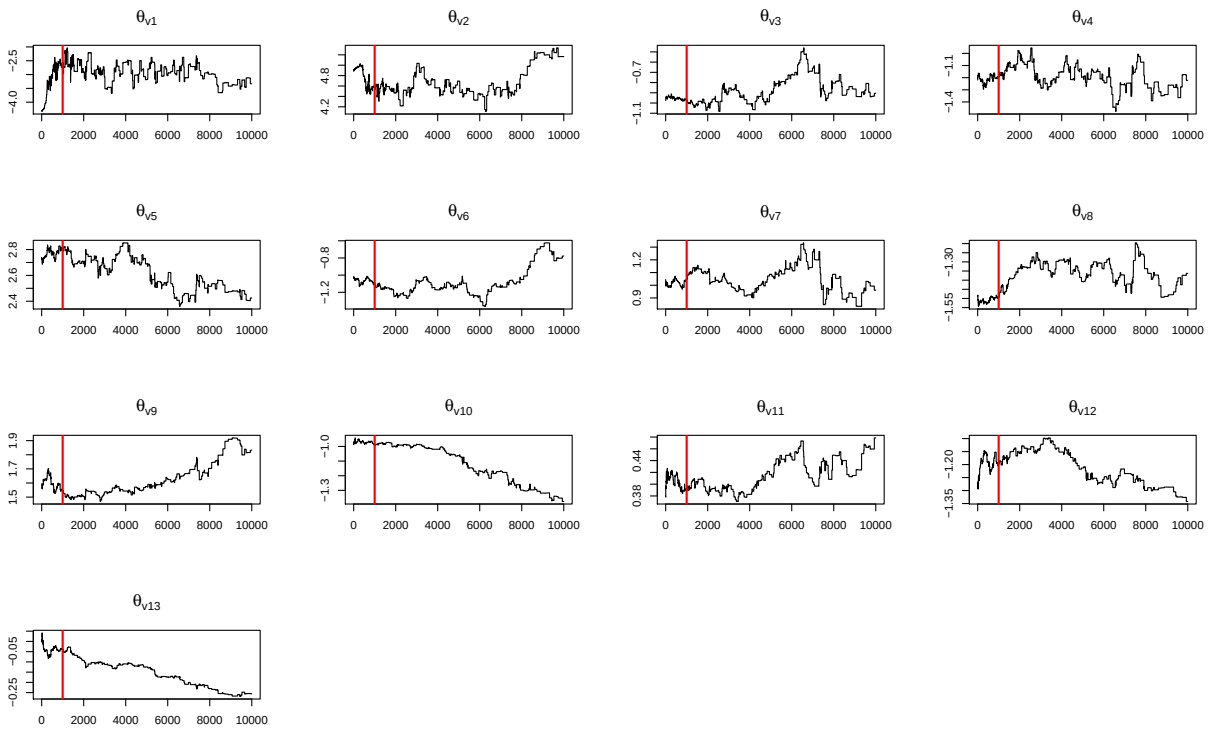


Figure 3.9: Trace plots of the parameter chains associated with the elements of $\boldsymbol{\theta}_V$ for estimating $v_1(x)$ via the DRAM procedure described in Section 3.2.1.

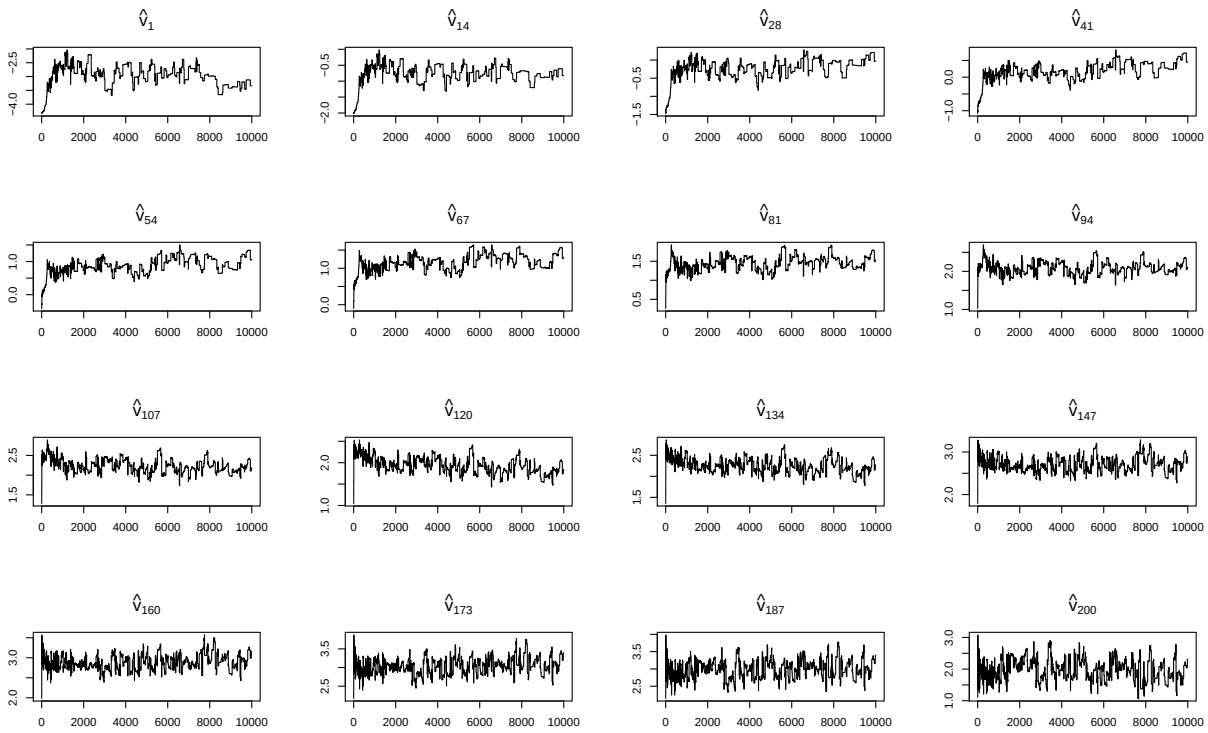


Figure 3.10: Selected pointwise trace plots for the estimate $\hat{v}_1(x)$ using the DRAM procedure described in Section 3.2.1.

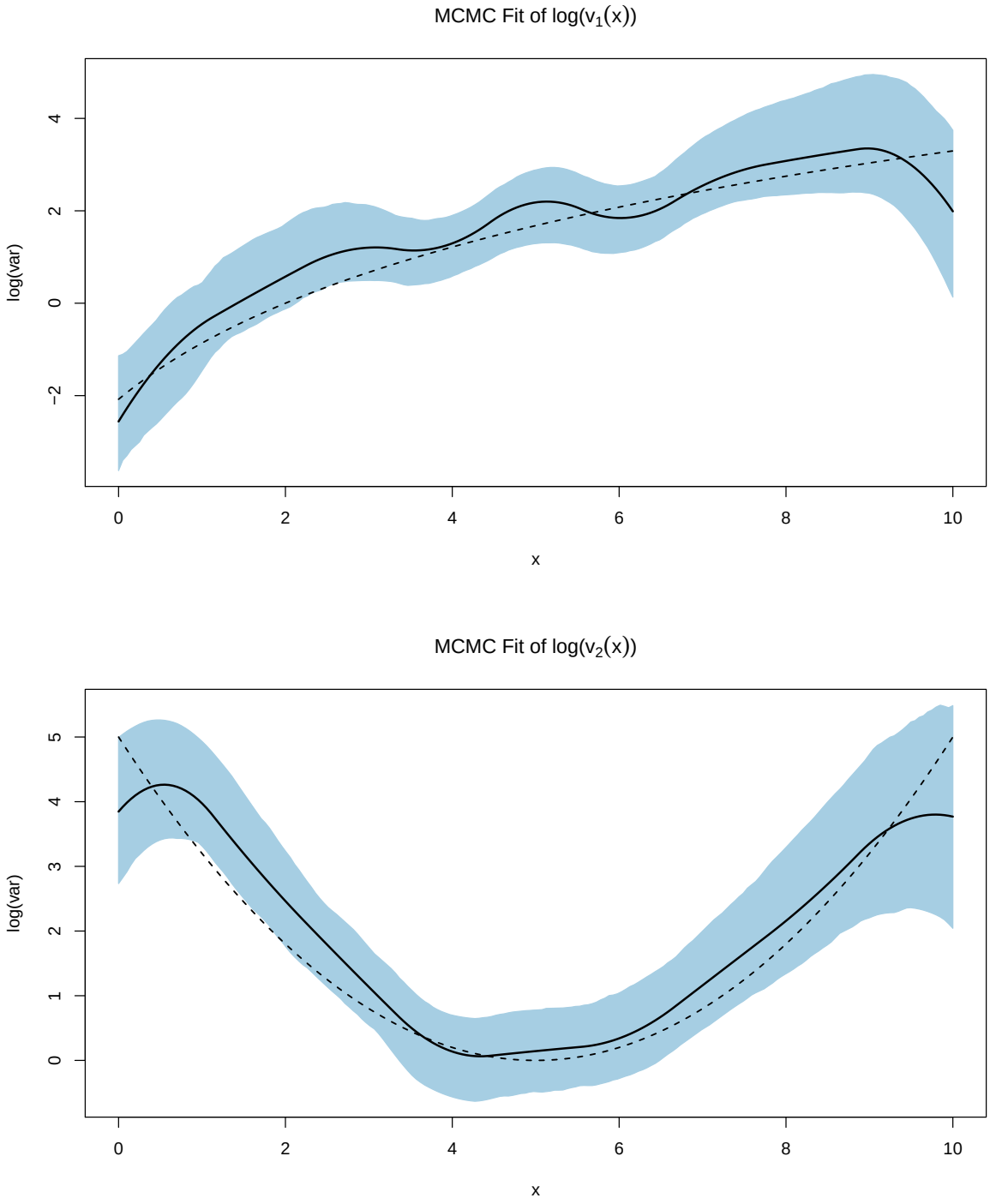


Figure 3.11: Estimated $\log(v(x))$ functions for single curve example after 50000 iterations with a 5000 iteration burn-in.

3.2.3 Vertically Shifted Curves with Common Variance Structure

As mentioned previously, a significant advantage of the mixed model framework for semiparametric regression is the ease with which one is able to extend models to include interaction effects. For heteroskedastic semiparametric regression, the examples presented up to now have described only single mean and variance curve scenarios. While this is quite common, there are many instances where multiple curves relating to additional covariates or interaction effects are of interest. In this section we continue our simulation study of the properties of the MCMC procedure when describing more complex multiple curve structures. For the following examples, all hyperparameters for fixed effect variances are set at $\sigma_\beta^2 = \sigma_\delta^2 = 10^5$. The rate and scale hyper parameters for the Inverse Gamma priors on the random effect variance terms are each set at 10^{-5}. We use a truncated quadratic spline basis evaluated at a set of 15 knots at equally-spaced quantiles ranging from 5% to 95% in each example at both the levels of the heteroskedastic model.

The first example considers the case where we have two curves whose only apparent difference is a vertical shift by a constant over the entire range of $x \in [0, 10]$. The true mean curves correspond to $m_1(x) = -0.125(x-5)^3 + x$ and $m_2(x) = -0.125(x-5)^3 + x + 10$ (plotted in green and purple respectively). We generate 200 samples from each curve, shown in Figure 3.12. Both mean curves share a common variance function $v(x) = \exp\left(-(x-5)^2/5\right)$. If we let $S_i = 10$ for data originating from $m_2(x)$ and $S_i = 0$ otherwise, we can model this example by including an additional fixed effect term to the model found in (35), taking the form

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 S_i + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+^2 + \epsilon_i \\ \log(\sigma_i^2) &= \delta_0 + \delta_1 x_i + \delta_2 x_i^2 + \sum_{k=1}^{K_V} c_k (x_i - \kappa_k^V)_+^2. \end{aligned} \tag{36}$$

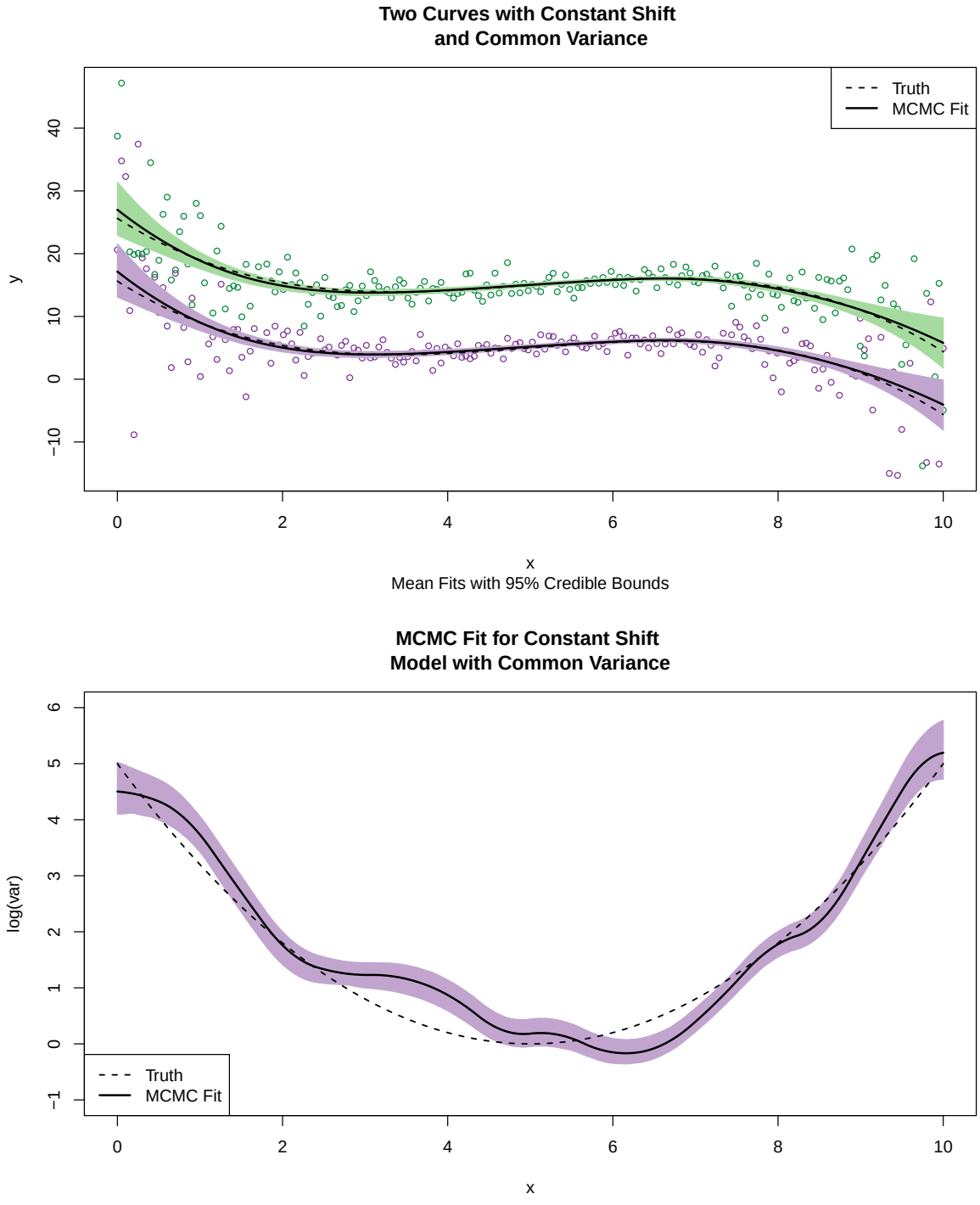


Figure 3.12: MCMC estimates of $m(x)$ (top) and $\log(v(x))$ for the parametric shift model found in (36). The MCMC procedure was run for $10,000$ iterations with a burn-in of 1000.

Figure 3.12 shows the resulting fits at both the mean and variance level for this model. As with the single curve examples detailed previously, our heteroskedastic model better portrays the observed uncertainty associated with the mean level data than a constant-errors version. The uncertainty estimates represented by the 95% credible bounds on the estimates of $m(x)$ pick up the increased variation towards the tails of the region of interest. The bottom panel of Figure 3.12 details the fit of the latent $\log(v(x))$ function. Here we observe similar behavior as with the single curve examples. The estimate of $\log(v(x))$ takes the general quadratic shape of the true $\log(v(x))$ but appears to miss the mark at several points across $[0, 10]$.

3.2.4 Interaction Model with Common Variance Structure

Our second example considers a situation where the interaction structure between two curves is more complex than a simple vertical translation. Data are simulated from mean curves $m_1(x) = \exp(-x^2/12)$ and $m_2(x) = \exp(-x^2/(12+S))$. For the data presented in Figure 3.13, 200 data points are simulated from $m_1(x)$ and $m_2(x)$ and $S = 8$. Both curves follow a common variance function $v(x) = (x/100)^2$. Let $S_i = 8$ for y_i simulated under $m_2(x)$ and 0 otherwise. Here we fit these data using the parametric-by-nonparametric interaction model

$$\begin{aligned} y_i &= f_1(x_i) + S_i f_2(x_i) + \epsilon_i \\ \log(\sigma_i^2) &= g(x_i). \end{aligned} \tag{37}$$

Using our truncated quadratic basis, this model expands to

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 S_i + \beta_4 x_i S_i + \beta_5 x_i^2 S_i \\ &\quad + \sum_{j=1}^{K} b_{1j}(x_i - \kappa_j)_+^2 + \sum_{j=1}^{K} S_i b_{2j}(x_i - \kappa_j)_+^2 + \epsilon_i \\ \log(\sigma_i^2) &= \delta_0 + \delta_1 x_i + \delta_2 x_i^2 + \sum_{k=1}^{K_V} c_k(x_i - \kappa_k)_+^2. \end{aligned} \tag{38}$$

The inclusion of the parametric-by-nonparametric interaction term, $S_i f_2(x_i)$, is handled through the additional random effect terms, $\{b_{2j}\}$, which have their own distinct variance term $\sigma^2_{b_2}$. Even without knowledge that a single variance function was used in the simulation, visual inspection of the data would support that while the underlying variance structure is most definitely non-constant, there is no evidence for dramatic differences across mean functions.

Figure 3.13 contains the results of our hybrid Gibbs-DRAM MCMC procedure for both the mean and variance level. The results are much the same as our previous single curve and multiple curve examples. Even under a more complicated mean structure that included two distinct random effect variance terms, our model accurately fits the data. The interaction term captures the nonlinear relationship between the shift covariate S and the underlying mean function while reflecting the increase in variation as x increases. As observed before, the underlying fit of $\log(v(x))$ closely tracks the true function but suffers from some fit difficulties particularly for x in the region $[6, 8]$.

3.2.5 Interaction Model with Dampening Variance

The final simulated example we consider for our heteroskedastic semiparametric regression MCMC procedure extends the interaction model described in (38) to include an interaction term at the variance level. As with before, 200 data points are simulated from $m_1(x) = \exp\left(-x^2/(12)\right)$ and $m_2(x) = \exp\left(-x^2/(12+S)\right)$ with $S = 8$. The variance function associated with $m_1(x)$ is $v_1(x) = (x/100)^2$, the same as the previous examples. For values generated from $m_2(x)$, we use a damped version of $v_1(x)$ defined as $v_2(x) = 0.05(x/100)^2$. Let D_i be an indicator variable that takes values 1 for responses generated from $m_2(x)$ and 0 otherwise. Using our penalized spline notation, this model is written as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 S_i + \beta_4 x_i S_i + \beta_5 x_i^2 S_i$$

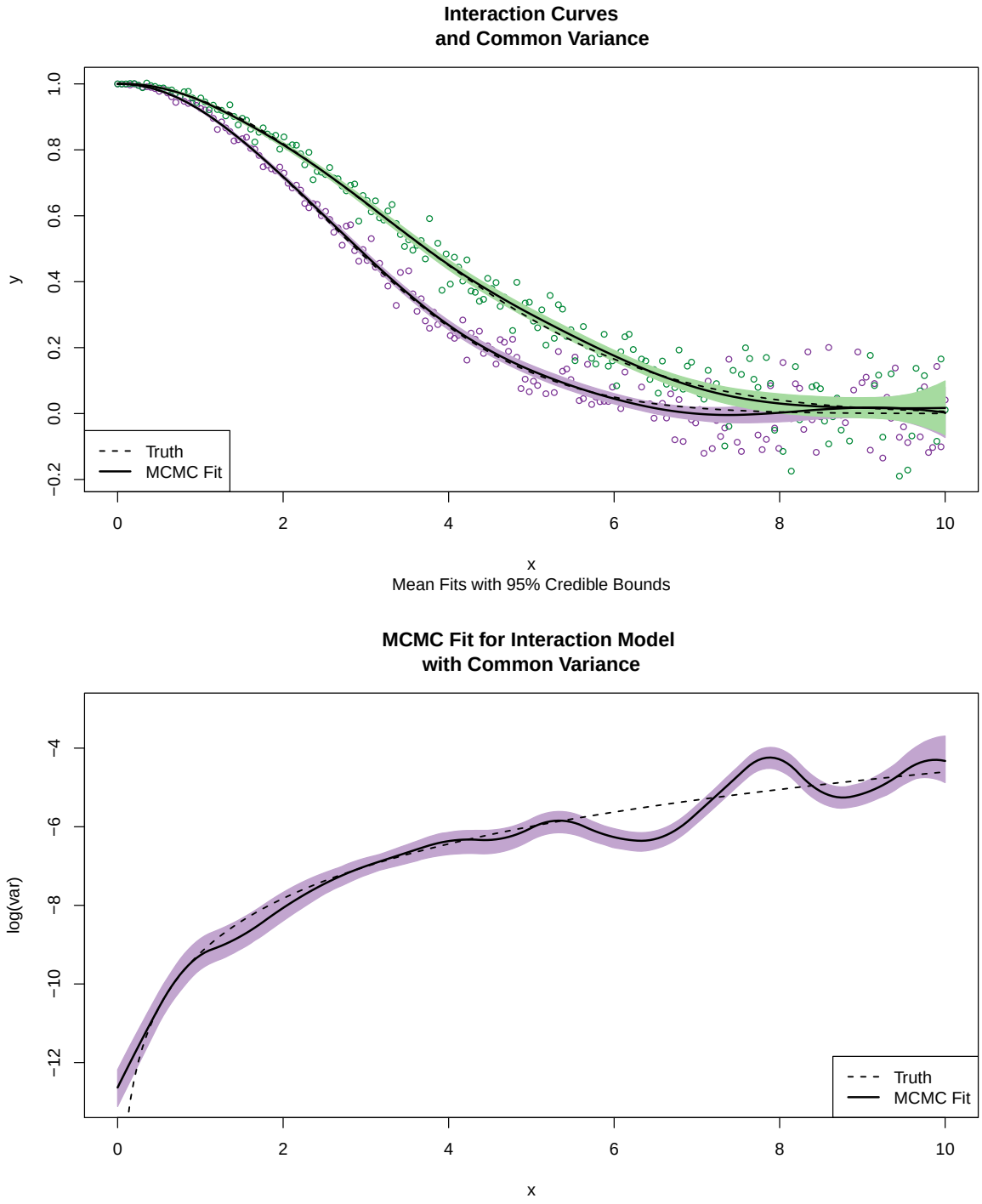


Figure 3.13: MCMC estimates of the mean and log variance function for the parametric-by-nonparametric interaction model with constant variance structure described in (38). The MCMC procedure was ran for $10,000$ iterations with a burn-in of 1000

$$+ \sum_{j=1}^{K} b_{1j}(x_i - \kappa_j)_+^2 + \sum_{j=1}^{K} S_i b_{2j}(x_i - \kappa_j)_+^2 + \epsilon_i$$

$$\log(\sigma_i^2) = \delta_0 + \delta_1 x_i + \delta_2 x_i^2 + \delta_3 D_i + \delta_4 x_i D_i + \delta_5 x_i^2 D_i$$
$$+ \sum_{k=1}^{K_V} c_{1k}(x_i - \kappa_{V_k})_+^2 + \sum_{k=1}^{K_V} D_i c_{2k}(x_i - \kappa_{V_k})_+^2. \quad (39)$$

As with the interaction model with common variance structure from (38), the parametric-by-nonparametric interaction terms are handled by the inclusion of additional random effect terms with distinct variance parameters.

Figure 3.14 shows the results from our MCMC procedure for this model. As with previous iterations, the model accurately fits the nonlinear interaction structure of the mean data while tracking the general structure of the latent log variance curves, but suffers from lack of fit issues in higher variance regions. Since this is the most complicated variance structure presented in our simulation study, it is not surprising that the estimates of $\log(v_1(x))$ and $\log(v_2(x))$ suffer the greatest difficulty we have observed so far.

The prevailing conclusion for both the single and multiple curve simulation examples for the heteroskedastic semiparametric problem is that our method performs quite well for the modeling of observed mean level data. However, while our model is able to better describe the underlying variance structure, some issues regarding coverage and lack of fit seem to plague our estimates of $\log(v(x))$, especially for more complicated variance structures. In the single curve examples, this lack of fit was partially resolved through an increase in the overall runtime of the MCMC procedure (10000 to 50000). For the more complicated models, particularly those that include interaction terms at the mean level, a 50000 length run seemed to yield only marginal improvements in the log variance estimates. In both the single curve and more complicated cases, our DRAM method seems to suffer from convergence issues. We conjecture that much of the difficulty associated with finding appropriate $\boldsymbol{\theta}_V$ comes from this

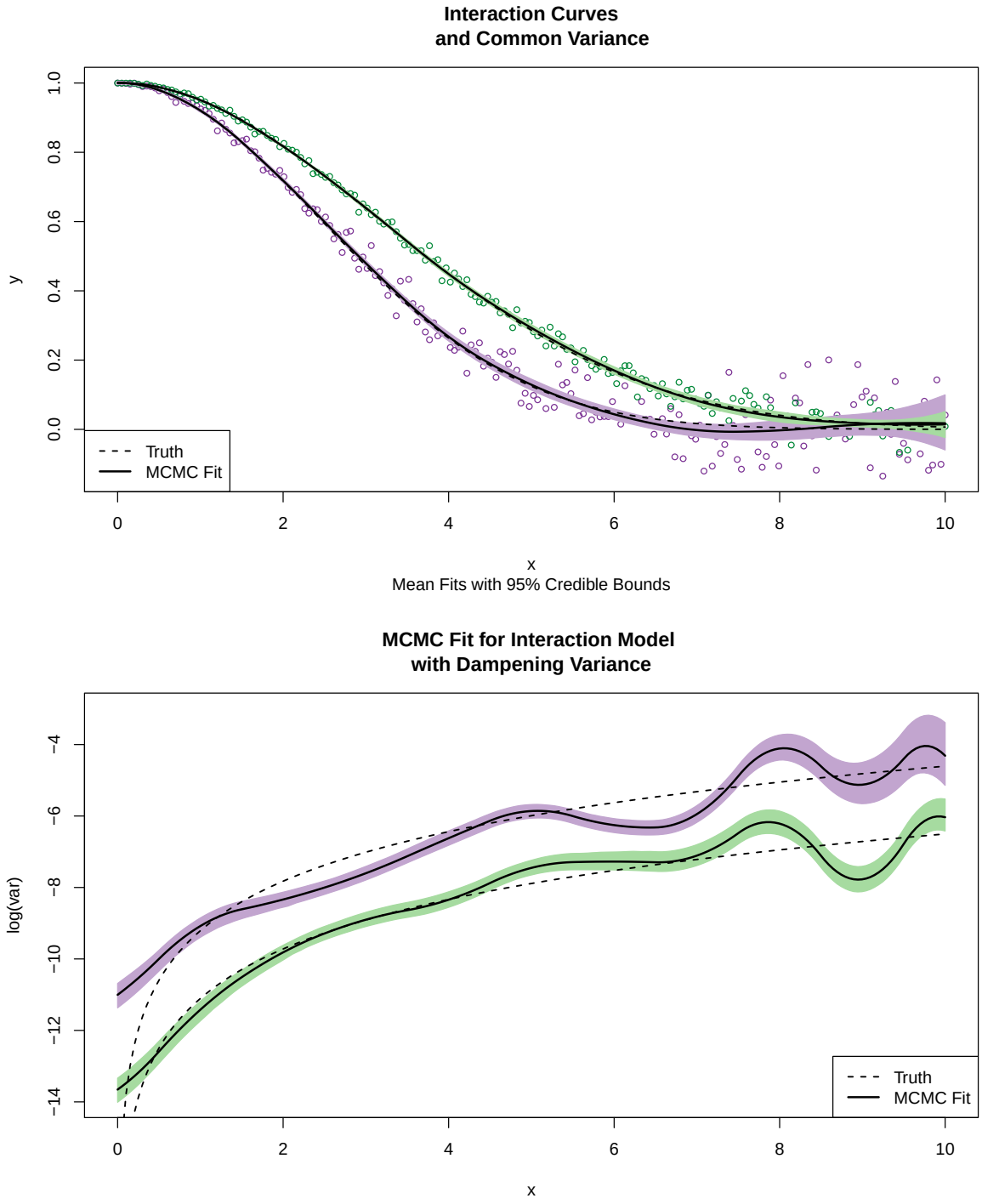


Figure 3.14: MCMC estimates of the mean and log variance function for the parametric-by-nonparametric interaction model with variance dampening described in (38). The MCMC procedure was run for $10,000$ iterations with a burn-in of 1000.

latent variable problem. It seems that for data at a given sample size, there is a class of similar curves that could be responsible for the observed variation. As sample size increases, this class gets increasingly smaller since we have more empirical variation observations to use in lieu of direct observations of the true variance function. This latent structure combined with computational difficulties associated with sampling $\boldsymbol{\theta}_V \mid \cdot$ potentially limits the quality of variance function estimates. Additionally, the associated computational issues of traversing the $\boldsymbol{\theta}_V$ parameter space can have impact on the overall fit of the log variance estimates. As discussed previously, investigating solutions for this problem in the form of additional DRAM modification, alternative basis functions, and parameter expansion are priorities for future work.

3.3 Joint Mean Variance Semiparametric Model via Hybrid Gibbs Sampling

As with the heteroskedastic errors model, the hierarchical nature of the joint mean-variance model detailed in Section 2.3 lends itself to a Bayesian framework. For the sake of clarity we first present work based on single pairs of radial mean and radial variance data. We primarily study the implementation of a computational scheme related to the method found in Section 3.2 for the case where the radial variance functional relationship, $h(s, \sigma^2)$ is modeled solely with fixed effects. Following this, we present the MCMC extension to the general model where all three levels can treated as generic combinations of parametric, nonparametric, and interaction terms.

3.3.1 Fixed Effect Radial Variance Model

Let $\mathbf{y} = \left(\bar{I}_1, \bar{I}_2, \ldots, \bar{I}_N\right)^T$, $\mathbf{w} = (\log(t_1^2), \log(t_2^2), \ldots, \log(t_N^2))^T$, and $\mathbf{v} = (\log(\sigma_1^2), \log(\sigma_2^2), \ldots, \log(\sigma_N^2))^T$. Define $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$ as the spatial covariate corresponding to radial distance. For convenience, the functional relationships at the mean and model variance level, $f(x)$ and $g(x)$ from (18), will be handled through a truncated quadratic spline basis expansion akin to previous work. Equation (40) describes the element-wise joint mean-variance model with radial variance fixed effects.

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sum_{k=1}^{K} b_k \left(x_i - \kappa_k\right)_+^2 + \epsilon_i \\
b_k &\sim \mathcal{N}\left(0, \sigma_b^2\right) \qquad \forall\, k = 1, \ldots K \\
\epsilon_i &\sim \mathcal{N}\left(0, \sigma_i^2\right) \\
v_i &= \delta_0 + \delta_1 x_i + \delta_2 x_i^2 + \sum_{k=1}^{K_V} c_k \left(x_i - \kappa_k^V\right)_+^2 \\
c_k &\sim \mathcal{N}\left(0, \sigma_c^2\right) \qquad \forall\, k = 1, \ldots, K_V
\end{aligned}
$$

$$
\begin{aligned}
w_i &= \eta_0 + \eta_1 \log(x_i) + \eta_2 v_i + u_i \\
u_i &\sim \mathcal{N}\left(0, \sigma_u^2\right).
\end{aligned} \tag{40}
$$

Written in matrix form, this model is

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \epsilon \\
\mathbf{b} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_b^2 \mathcal{I}_K\right) \\
\epsilon &\sim \mathcal{N}\left(0, \boldsymbol{\Sigma}\right) \\
\mathbf{v} &= \mathbf{X}_V\boldsymbol{\delta} + \mathbf{Z}_V\mathbf{c} \\
\mathbf{c} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_c^2 \mathcal{I}_{K_V}\right) \\
\mathbf{w} &= \mathbf{X}_W\boldsymbol{\eta} + \mathbf{u} \\
\mathbf{u} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_u^2 \mathcal{I}_N\right).
\end{aligned} \tag{41}
$$

As with before, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and $\boldsymbol{\eta}$ are vectors of fixed effect parameters and $\mathbf{b}$ and $\mathbf{c}$ are vectors of random effects. The matrix $\boldsymbol{\Sigma}$ correspond to a diagonal covariance matrix where $\boldsymbol{\Sigma}_{ii} = \sigma_i^2$. For notational convenience, let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{b})^T$, $\boldsymbol{\theta}_V = (\boldsymbol{\delta}, \mathbf{c})^T$, $\mathbf{C} = [\mathbf{X}, \mathbf{Z}]$, and $\mathbf{C}_V = [\mathbf{X}_V, \mathbf{Z}_V]$.

Based on the similarity to the heteroskedastic model, we choose conjugate priors with distributions

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\beta^2 \mathcal{I}_3\right) \\
\boldsymbol{\delta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\delta^2 \mathcal{I}_3\right) \\
\boldsymbol{\eta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\eta^2 \mathcal{I}_3\right) \\
\sigma_b^2 &\sim \mathcal{IG}\left(A_b, B_b\right) \\
\sigma_c^2 &\sim \mathcal{IG}\left(A_c, B_c\right) \\
\sigma_u^2 &\sim \mathcal{IG}\left(A_u, B_u\right),
\end{aligned} \tag{42}
$$

where σ_β^2, σ_δ^2, σ_η^2, A_b, B_b, A_c, B_c, A_u, and B_u are fixed hyperparamters. Under these priors, the posterior conditional distributions for the parameters of interest are shown in Equation (43).

$$
\begin{aligned}
\boldsymbol{\theta} \mid \cdot &\sim \mathcal{N}\left(\mathbf{M}\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}, \mathbf{M}\right) \text{ where } \mathbf{M} = \left(\boldsymbol{\Sigma}_\theta{}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}\right)^{-1} \\
\boldsymbol{\eta} \mid \cdot &\sim \mathcal{N}\left(\frac{1}{\sigma_u^2}\mathbf{M}_W\mathbf{X}_W^T\mathbf{w}, \mathbf{M}_W\right) \text{ where } \mathbf{M}_W = \left(\boldsymbol{\Sigma}_\eta{}^{-1} + \frac{1}{\sigma_u^2}\mathbf{X}^T\mathbf{X}\right)^{-1} \\
\sigma_b^2 \mid \cdot &\sim \mathcal{IG}\left(A_b + \frac{K}{2}, B_b + \frac{\|\mathbf{b}\|^2}{2}\right) \\
\sigma_c^2 \mid \cdot &\sim \mathcal{IG}\left(A_c + \frac{K_V}{2}, B_c + \frac{\|\mathbf{c}\|^2}{2}\right) \\
p(\boldsymbol{\theta}_V \mid \cdot) &\propto \exp\Big[-\frac{1}{2}\Big\{ \sum_{i=1}^{N} \mathbf{C_{V_i}}{}^T\theta_{\mathbf{V}} + \sum_{i=1}^{N}\left(Y_i - \mathbf{C_i}{}^T\theta\right)^2 \exp(-\mathbf{C_{V_i}}{}^T\theta_{\mathbf{V}}) \\
&\qquad + \theta_{\mathbf{V}}{}^T\boldsymbol{\Sigma}_{\theta_{\mathbf{V}}}{}^{-1}\theta_{\mathbf{V}} + \frac{1}{\sigma^2}\|\mathbf{w} - \mathbf{X}_W\boldsymbol{\eta}\|^2 \Big\}\Big]. \\
\sigma_u^2 \mid \cdot &\sim \mathcal{IG}\left(A_u s + \frac{N}{2}, B_u + \frac{\|\mathbf{w} - \mathbf{X}_W\boldsymbol{\eta}\|^2}{2}\right)
\end{aligned}
\tag{43}
$$

The posterior conditional distributions presented here are quite similar to those found in the heteroskedastic semiparametric regression model with two main departures. The first departure is the addition of posterior conditionals for $\boldsymbol{\eta}$ and σ_u^2. Since these have known forms, it is a simple matter to include them in our Gibbs sampling procedure as outlined previously. The second departure comes in the additional term in the non-conjugate distribution for $\boldsymbol{\theta}_V \mid \cdot$. The proportional form of $p(\boldsymbol{\theta}_V \mid \cdot)$ here includes the term $\|\mathbf{w} - \mathbf{X}_W\boldsymbol{\eta}\|^2/\sigma_u^2$. Given the fixed effect structure detailed in (40), $\mathbf{v}$ is a covariate which embeds a relationship with $\boldsymbol{\theta}_V$ in the matrix $\mathbf{X}$. As such, this term is included in the derivation of $p(\boldsymbol{\theta}_V \mid \cdot)$.

To sample from these posterior conditionals, we propose to use the same hybrid Gibbs sampler as the heteroskedsatic semiparametric regression problem detailed in Section 3.2. A two-stage DRAM step is used to sample from the posterior conditional distribution of $\boldsymbol{\theta}_V \mid \cdot$

while the other parameters of interest are sampled directly from their known forms. For now we use the same tuning parameters and Gaussian proposals for the DRAM step as previous work.

3.3.2 Radial Fixed Effect Simulation Example

To test this joint methodology we consider a simulated data example. Rather than attempting to simulate the physical process that generates radial data of this type, we simulated a heteroskedastic "radial mean" curve with known mean and model variance functions similar to the simulation examples from Section 3.2.2. Then our "radial variance" data is simulated with a known form based on spatial covariate x and the true log model variance v. The data for this example are generated according to

$$\begin{aligned} m(x) &= 25(e^{-(x-1)^2/10} + 3) \\ \log\left(\sigma^2(x)\right) &= \log\left(\frac{2x^2+10}{20}\right) \\ \log\left(t^2(x,\sigma^2(x))\right) &= 10 - 4\log(x) + \log\left(\sigma^2(x)\right), \end{aligned} \tag{44}$$

where $m(x)$ is the true "radial mean" function with true model variance $\sigma^2(x)$. The function $\log\left(t^2(x,\sigma^2(x))\right)$ represents true log "radial variance" function here. For notational simplicity, let $w(x,\sigma^2) = \log\left(t^2(x,\sigma^2(x))\right)$ and $v(x) = \log\left(\sigma^2(x)\right)$. The observed data in this example consists of $N = 400$ response pairs $\{y_i, w_i\}_{i=1}^N$ generated from evenly spaced x values across the interval $[1, 10]$. Gaussian noise with variance $\sigma_u^2 = 0.25$ is added to generate the observed w_i values. Figure 3.15 shows the simulated "observed" data.

The data is fit with the hierarchical model described in (40). The mean and model variance levels consist of nonparametric fits using a truncated quadratic spline basis ($p = 2$) of $K = K_V = 10$ quantile spaced knots across the range of x. The radial variance level is fit via a fixed effects model with both $\log(x)$ and v as covariates. All Inverse Gamma hyper

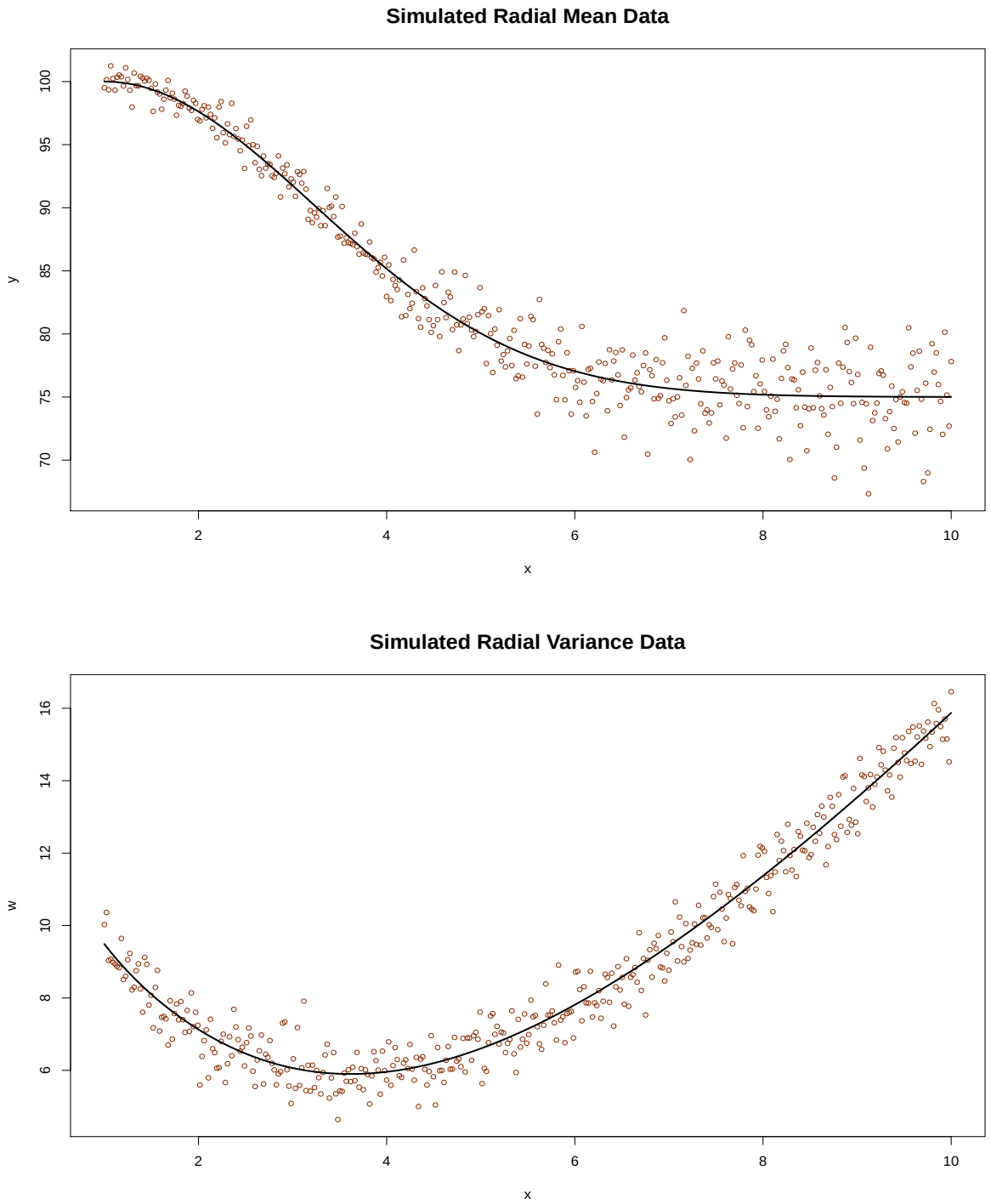


Figure 3.15: Simulated data meant to mimic the radial mean and variance data arising from SAXS experiments. The top panel corresponds to radial mean responses simulated from $m(x) = 25(e^{-(x-1)^2/10} + 3)$ with model log variance function $v(x) = \log\left(\frac{2x^2+10}{20}\right)$. The bottom panel corresponds to the radial variance data simulated according to $10 - 4\log(x) + v(x) + u$ where $u \sim \mathcal{N}(0, 0.25)$.

parameters are set to 10^{-8}. The variance hyperparameters for the fixed effect terms are $\sigma_\beta^2 = \sigma_\delta^2 = \sigma_\eta^2 = 10^5$.

Figure 3.16 contains the result of a 50000 iteration run of our MCMC procedure with a burn-in of 5000. Both radial mean and radial variance data are adequately fit by the model. The shaded regions correspond to 95% pointwise credible bounds for the smooth fits and reflect the variance structure of both response sets. The heteroskedastic behavior observed in the radial mean data is reflected by the increase in the bounds for larger values of x while the constant error structure of the radial variance data is maintained. Table 3.1 contains the MCMC estimates of the parameters used to generate the radial variance data. All estimates are very close to their true values. There is indication of slight coverage issues with the credible bounds associated with the parameter chains however. For this particular run, the credible bounds for η_1, η_2, and σ^2 barely miss the true value, while still being quite close. Most likely this has to do with the fact that these estimates are from a model where one of the covariates is the latent model variance v_i. Discrepancy in between the true model variance and the model estimate could result in the parameters of the radial variance level having different interpretability than before. In essence, they would be modeling a different covariate set than the one that generated the data. More detailed investigation of the implications of including the log model variance as a covariate in these models is an interesting topic for future work.

Table 3.1: Parameter estimates for simulated radial variance data under fixed effect structure. The MCMC procedure was ran of 50000 iterations.

	Estimate	2.5%	97.5%
η_0	9.45	8.57	10.17
η_1	-3.22	-3.98	-2.20
η_2	0.39	-0.27	0.84
σ_u^2	0.31	0.26	0.35

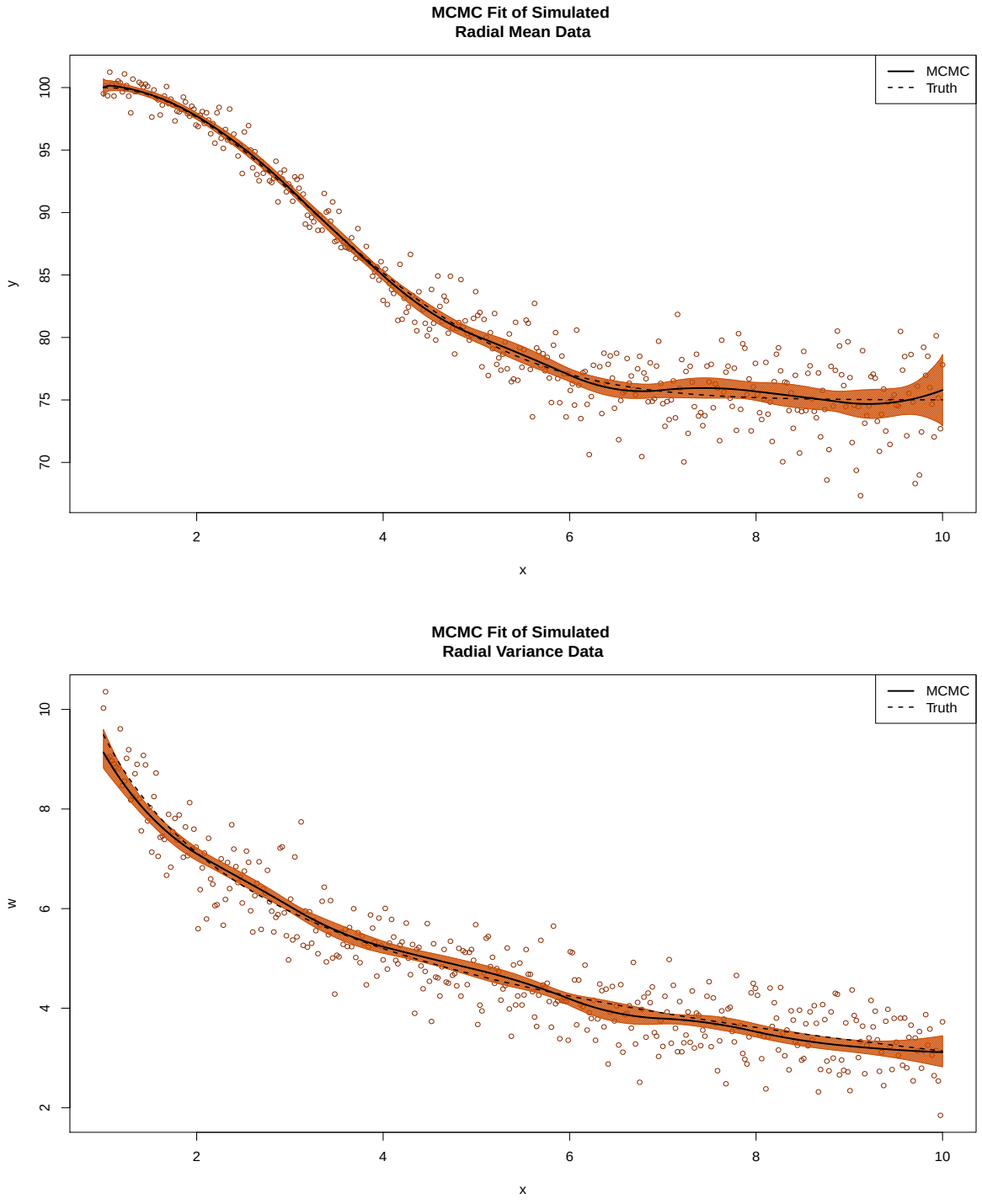


Figure 3.16: Estimates of the underlying radial mean and radial variance functions for the radial fixed effect simulation example. The MCMC procedure was ran for 50000 iterations with a burn-in of 5000

The top panel of Figure 3.16 compares the estimated log model variance function, $\hat{v}(x)$, with the true function. For comparison, the bottom panel of Figure 3.16 contains the $\hat{v}(x)$ from a 50000 iteration run of a heteroskedastic semiparametric regression model, using the same set up as our joint model where appropriate. As discussed in previous sections, precise fitting of the log model variance function used to generate the data can be difficult due to lack of direct observations. The MCMC estimates of $\hat{v}(x)$ in both cases appear to perform quite well. The only significant difference between the methods with regards to variance function estimate was observed for chains with smaller run-lengths. Figure 3.18 displays both fits resulting from a 10000 run iteration. The "short run" estimates of $v(x)$ suggest that including the radial variance structure may help the DRAM step traverse the multivariate posterior conditional $\boldsymbol{\theta}_V \mid \cdot$ and reach acceptable estimates faster than the alternative. It is unclear how much this improvement is contingent on the proper specification of the radial variance model. Ideally, improper specification of the radial variance model would result in MCMC procedure placing more weight on the inference drawn from the residuals of the radial mean level. Investigating this relationship is an interesting subject for future investigations.

The 50000 iteration run of the joint mean-variance MCMC procedure took approximately 10 minutes on a standard MacBook Pro laptop with a 2.3 GHZ Intel Core i5 processor and 4GB of RAM. Figure 3.19 consists of the element-wise trace plots of $\boldsymbol{\theta}_V$. The DRAM step had an overall proposal acceptance rate of approximately 4% for the length of this run, comparable to similar length runs of the heteroskedastic semiparametric regression DRAM method. Improved tuning of this sampling step may lead to faster convergence of the $\boldsymbol{\theta}_V$ chain, reducing the overall number of iterations needed.

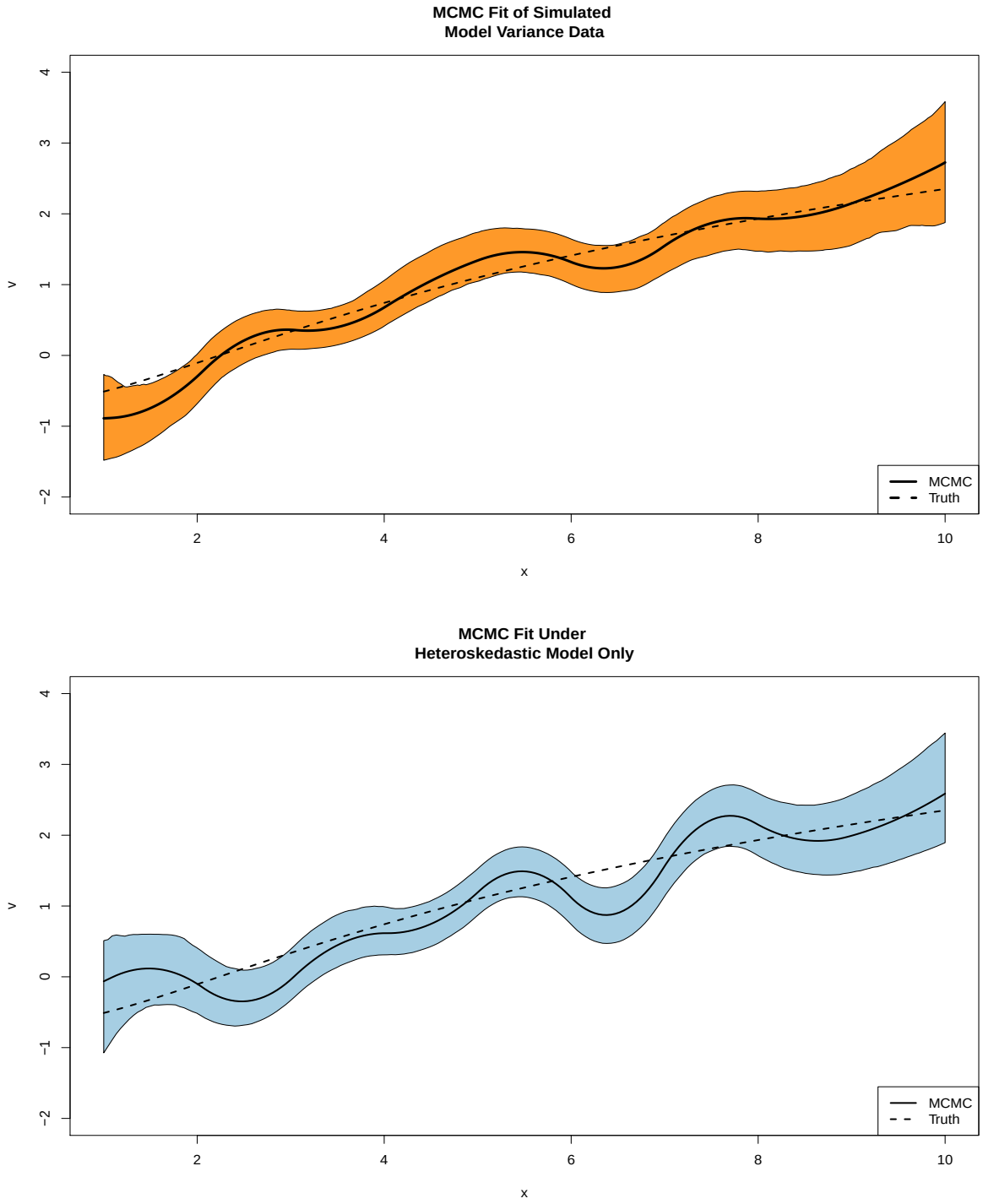


Figure 3.17: Estimates of the log model variance function for the radial fixed effect simulation example. The top panel is from the MCMC procedure for joint mean-variance data while the bottom panel corresponds to a heteroskedastic semiparametric regression model that ignores the radial variance data. Both MCMC procedures were ran for 50000 iterations with a burn-in of 5000.

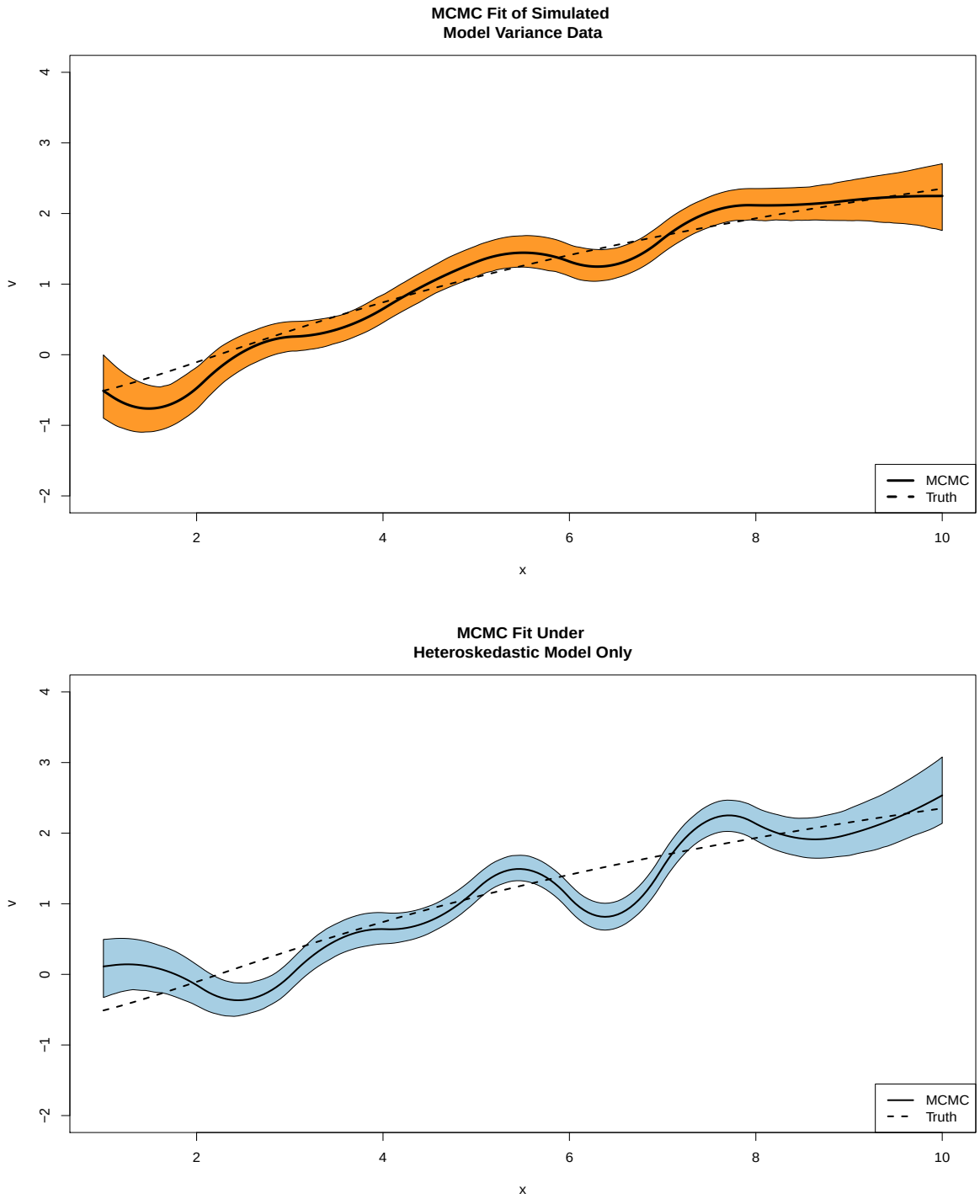


Figure 3.18: Log model variance estimates as shown in Figure 3.17 for a 10000 iteration run.

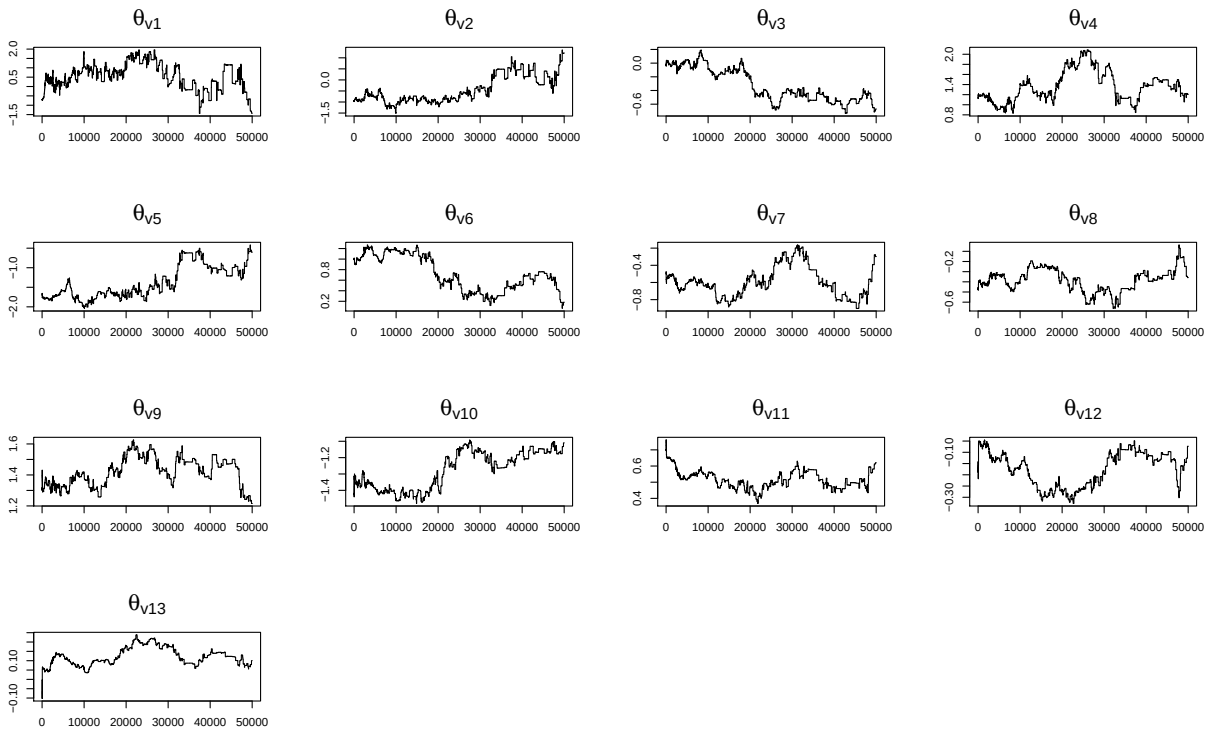


Figure 3.19: Trace plots of the $\boldsymbol{\theta}_V$ parameter chain from a 50000 iteration run of the MCMC procedure for the radial fixed effect simulation example.

3.3.3 General Joint Model

It is relatively simple to extend the model described in (41) to allow for more complicated parametric and nonparametric terms at all three levels. In general terms, the joint mean-variance model is written as:

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{b}_1 + \mathbf{Z}_2\mathbf{b}_2 + \cdots + \mathbf{Z}_L\mathbf{b}_L + \epsilon \\
\mathbf{b}_l &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{b_l}\mathcal{I}_{K_l}\right) \ \ \forall\, l = 1, \ldots, L \\
\epsilon &\sim \mathcal{N}\left(0, \boldsymbol{\Sigma}\right) \\
\mathbf{v} &= \mathbf{X}_V\boldsymbol{\delta} + \mathbf{Z}_{V_1}\mathbf{c}_1 + \mathbf{Z}_{V_2}\mathbf{c}_2 + \cdots + \mathbf{Z}_{V_M}\mathbf{c}_M \\
\mathbf{c}_m &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{c_m}\mathcal{I}_{K_{V_m}}\right) \ \ \forall\, m = 1, \ldots, M \\
\mathbf{w} &= \mathbf{X}_W\boldsymbol{\eta} + \mathbf{Z}_{W_1}\mathbf{d}_1 + \mathbf{Z}_{W_2}\mathbf{d}_2 + \cdots + \mathbf{Z}_{W_R}\mathbf{d}_R + \mathbf{u} \\
\mathbf{b}_r &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{b_\rho}\mathcal{I}_{K_{W_\rho}}\right) \ \ \forall\, \rho = 1, \ldots, R \\
\mathbf{u} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_u\mathcal{I}_N\right).
\end{aligned} \tag{45}$$

Recall that $\boldsymbol{\Sigma}$ is the diagonal covariance matrix with entries σ_i^2 and $\mathbf{v} = \left(\log(\sigma_1^2), \ldots, \log(\sigma_N^2)\right)^T$. Generic inclusion of mixed effects at all three levels of the model allow for both parametric and nonparametric terms to be specified. New random effects $\mathbf{d}_1$, $\mathbf{d}_2$, $\ldots$, $\mathbf{d}_R$ are included for the radial variance level. Define $\mathbf{C}_W = [\mathbf{X}_W, \mathbf{Z}_{W_1}, \ldots, \mathbf{Z}_{W_R}]$ and $\theta_{\mathbf{W}} = \left(\boldsymbol{\eta}^T, \mathbf{d}_1^T, \ldots, \mathbf{d}_R^T\right)^T$. The generalized conjugate prior structure is

$$\begin{aligned}
\boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_\beta\mathcal{I}_3\right) \\
\boldsymbol{\delta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_\delta\mathcal{I}_3\right) \\
\boldsymbol{\eta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_\eta.\mathcal{I}_3\right) \\
\sigma^2_{b_l} &\sim \mathcal{IG}\left(A_{b_l}, B_{b_l}\right) \ \ \forall l = 1, \ldots, L \\
\sigma^2_{c_m} &\sim \mathcal{IG}\left(A_{c_m}, B_{c_m}\right) \ \ \forall m = 1, \ldots, M
\end{aligned}$$

$$\sigma^2_{d_\rho} \sim \mathcal{IG}\left(A_{d_\rho}, B_{d_\rho}\right) \;\; \forall \rho = 1, \ldots, R$$
$$\sigma^2_u \sim \mathcal{IG}\left(A_u, B_u\right), \tag{46}$$

with fixed hyperparameters σ^2_β, σ^2_δ, σ^2_η, $\{A_{b_l}, B_{b_l}\}_{l=1}^L$, $\{A_{c_m}, B_{c_m}\}_{m=1}^M$, $\left\{A_{d_\rho}, B_{d_\rho}\right\}_{\rho=1}^R$, A_u, and B_u. The posterior conditional distributions for the parameters of interest are

$$\begin{aligned}
\boldsymbol{\theta} \mid \cdot &\sim \mathcal{N}\left(\mathbf{M}\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}, \mathbf{M}\right) \text{ where } \mathbf{M} = \left({\boldsymbol{\Sigma}_\theta}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}\right)^{-1} \\
\theta_{\mathbf{W}} \mid \cdot &\sim \mathcal{N}\left(\frac{1}{\sigma^2_u}\mathbf{M}_W\mathbf{C}_W^T\mathbf{w}, \mathbf{M}_W\right) \text{ where } \mathbf{M}_W = \left({\boldsymbol{\Sigma}_{\theta_\mathbf{W}}}^{-1} + \frac{1}{\sigma^2_u}\mathbf{C}_W^T\mathbf{C}_W\right)^{-1} \\
\sigma^2_{b_l} \mid \cdot &\sim \mathcal{IG}\left(A_{b_l} + \frac{K_l}{2}, B_{b_l} + \frac{\|\mathbf{b}_l\|^2}{2}\right) \;\; \forall\, l = 1, \ldots, L \\
\sigma^2_{c_m} \mid \cdot &\sim \mathcal{IG}\left(A_{c_m} + \frac{K_{V_m}}{2}, B_{c_m} + \frac{\|\mathbf{c}_m\|^2}{2}\right) \;\; \forall\, m = 1, \ldots, M \\
\sigma^2_{d_\rho} \mid \cdot &\sim \mathcal{IG}\left(A_{d_\rho} + \frac{K_{W_\rho}}{2}, B_{d_\rho} + \frac{\|\mathbf{d}_\rho\|^2}{2}\right) \;\; \forall\, \rho = 1, \ldots, R \\
p(\boldsymbol{\theta}_V \mid \cdot) &\propto \exp\Big[-\frac{1}{2}\Big\{\sum_{i=1}^N {\mathbf{C}_{\mathbf{V_i}}}^T\theta_\mathbf{V} + \sum_{i=1}^N \left(Y_i - {\mathbf{C_i}}^T\theta\right)^2 \exp(-{\mathbf{C}_{\mathbf{V_i}}}^T\theta_\mathbf{V}) \\
&\qquad + {\theta_\mathbf{V}}^T{\boldsymbol{\Sigma}_{\theta_\mathbf{V}}}^{-1}\theta_\mathbf{V} + \frac{1}{\sigma^2}\|\mathbf{w} - \mathbf{C}_W\theta_\mathbf{W}\|^2\Big\}\Big]. \\
\sigma^2_u \mid \cdot &\sim \mathcal{IG}\left(A_u + \frac{N}{2}, B_u + \frac{\|\mathbf{w} - \mathbf{C}_W\theta_\mathbf{W}\|^2}{2}\right).
\end{aligned} \tag{47}$$

The notation $\boldsymbol{\Sigma}_{\theta_\mathbf{W}}$ refers to the covariance matrix

$$\boldsymbol{\Sigma}_{\theta_\mathbf{W}} = \text{blockdiag}\left(\sigma^2_\eta\mathcal{I}_q, \sigma^2_{\mathbf{d}_1}\mathcal{I}_{W_1}, \ldots, \sigma^2_{\mathbf{d}_R}\mathcal{I}_{W_R}\right). \tag{48}$$

To sample from the posterior conditionals in (47), the same MCMC procedure as the radial variance fixed effect version is used. All parameters except for $\boldsymbol{\theta}_V$ are drawn directly from their known forms while a two-stage DRAM step is used to sample from the posterior conditional distribution of $\boldsymbol{\theta}_V \mid \cdot$.

3.3.4 General Joint Model Example

As with the preceding methodologies, the mixed model representation of the joint radial mean and variance model presented here allows for the flexible definition of complex structures at the radial mean, radial variance, and model variance levels. We now provide an illustrative simulation example of the general semiparametric methodology. Consider $N = 400$ radial mean observations as shown in Figure 3.20 (200 per mean function). The data are generated according to mean functions

$$\begin{aligned} m_1(x) &= 25(e^{-(x-1)^2/10} + 3) \\ m_2(x) &= m_1(x) + 15. \end{aligned} \tag{49}$$

The mean functions $m_1(x)$ and $m_2(x)$ share the log model variance function

$$v(x) = \log\left(\frac{2x^2 + 10}{20}\right). \tag{50}$$

The log radial variance responses are simulated according to the function

$$\begin{aligned} w(x) &= 10 - 4\log(x) + v(x) + 2v(x)\log(x) + u \\ u &\sim \mathcal{N}(0, 0.25). \end{aligned} \tag{51}$$

Cursory observation of the radial mean data suggests adding a parametric shift component to the nonparametric model used early. Element-wise, the radial mean is

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i + \beta_3 S_i + \sum_{k=1}^{K} b_k \left(x_i - \kappa_k\right)_+^2 + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}\left(0, \sigma_i^2\right). \end{aligned} \tag{52}$$

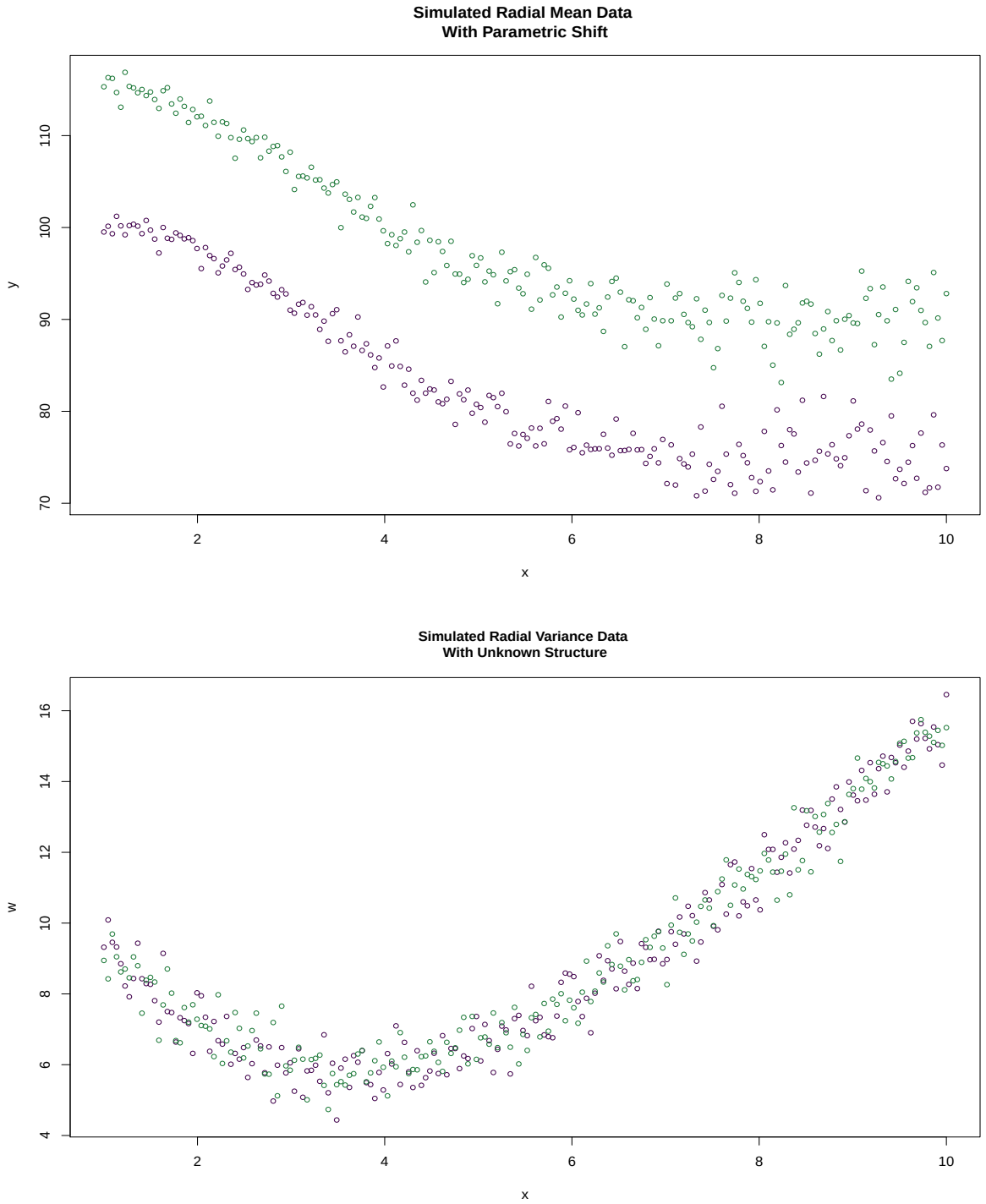


Figure 3.20: The top panel describes radial mean data simulated from $m_1(x) = 25(e^{-(x-1)^2/10} + 3)$ (purple) and $m_2(x) = m_1(x) + 15$ (green). Both mean functions have true log model variance $v(x) = \log\left(\frac{2x^2+10}{20}\right)$. The bottom panels shows the simulated radial variance data. Data here is simulated regardless of generating mean function following $w(x) = 10 - 4\log(x) + v(x) + 2v(x)\log(x) + u$ where $u \sim \mathcal{N}(0, 0.25)$.

The observed variation of the radial mean data suggests the same heteroskedastic behavior across mean curves. As such, we use the same truncated quadratic model as (40) for the model variance level.

For this example we choose to model the radial variance data as a general bivariate function of both x and σ^2. Element-wise, the model is

$$\begin{aligned} w_i &= \eta_0 + \sum_{\rho=1}^{R} d_\rho B_\rho\left(x_i, v_i\right) + u_i \\ \mathbf{d} &\sim \mathcal{N}\left(0, \sigma_d^2 \mathcal{I}_R\right) \\ u_i &\sim \mathcal{N}\left(0, \sigma_u^2\right) \end{aligned} \tag{53}$$

where the bivariate basis $\{B_\rho(x, v)\}_{\rho=1}^{R}$ corresponds to a set of thin-plate splines evaluated over a grid of $\mathrm{R} = 25$ equally spaced knots (Kammann and Wand, 2003). All hyperparameters are set to the same values as the previous example. The MCMC procedure was ran for 50000 iterations with a 5000 step burn-in time.

Figure 3.21 shows the results of this model for the simulated data. The mean level estimates, depicted in the top panel, pick up the parametric shift component while also appropriately reflecting the increased variation at larger values of x. The bottom panel shows the resulting fit from the bivariate thin-plate splines expansion for the radial variance data. Figure 3.22 shows the estimate of the log model variance function for both the joint model presented here as well as a standalone heteroskedastic semiparametric model. While the bounds associated with the heteroskedastic model are tighter in most places, this estimate suffers from the same difficulty of fit problem we have observed when estimating latent variance functions. The estimate associated with the joint model has better coverage properties of this value, albeit at the price of wider uncertainty bounds.

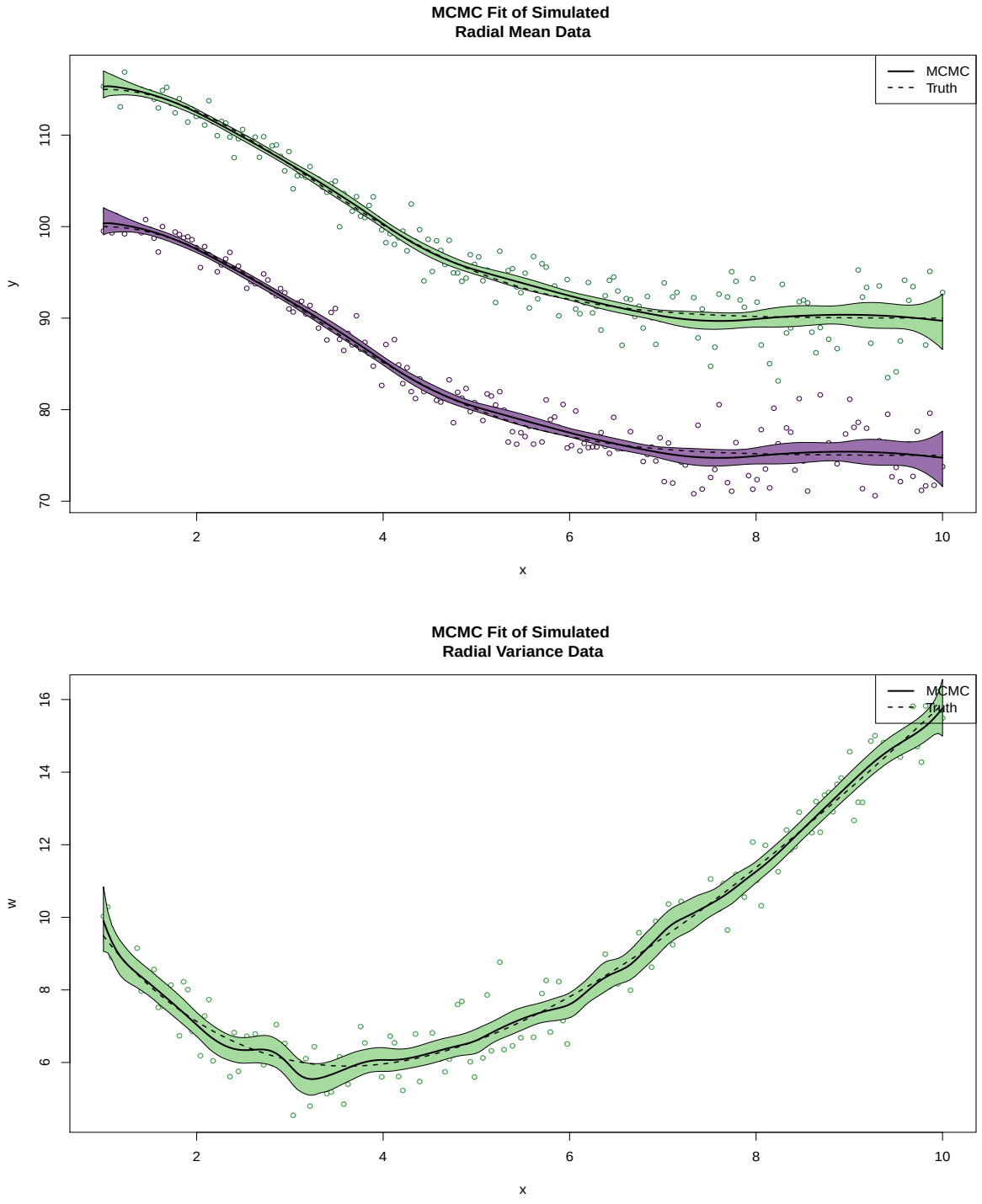


Figure 3.21: Estimated radial mean functions (top) and radial variance function (bottom) for the general joint model example. Shaded regions represent 95% pointwise credible intervals.

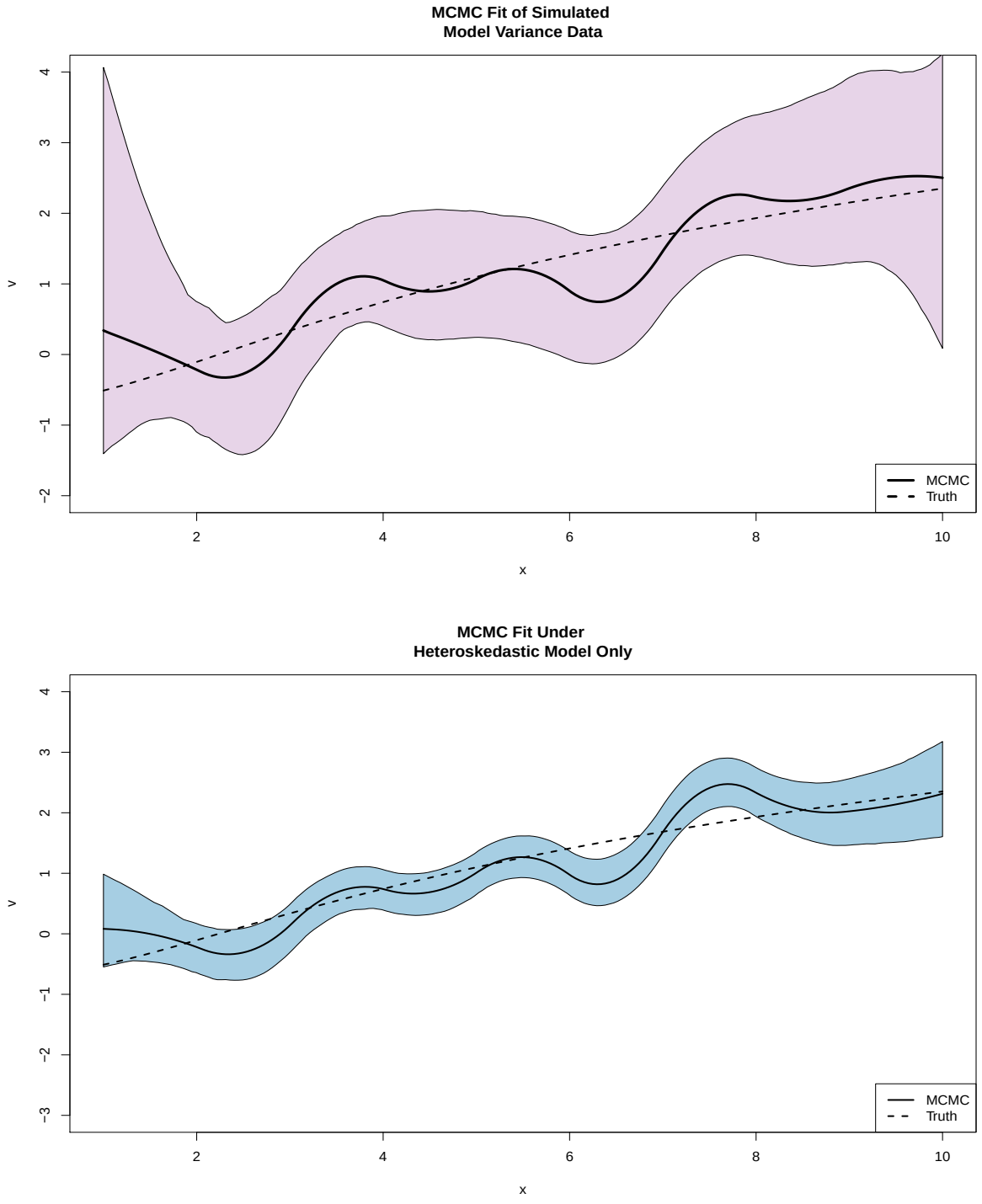


Figure 3.22: Estimated log model variance functions for both the joint mean-variance model (top) and the standard heteroskedastic model (bottom).Shaded regions represent 95% pointwise credible intervals.

The hierarchical model for radial mean and radial variance data and accompanying MCMC procedure described here have shown to be useful in describing the observed responses under various parametric and nonparametric forms. Also, there is anecdotal evidence that the structural link between radial and model variances can aid in estimating the true model variance function, particularly when additional effects are added to the radial mean level of the model. The simulation examples here were limited to cases that mimic behavior that could reasonably be seen in SAXS experimental data. This hierarchical model could be used for similar problems, where apart from observed heteroskedasticiy, we have an auxiliary response which is known to be related to the model variance. Investigate the properties of this model under more complicated scenarios as well as non-SAXS related problems that would benefit form this approach is an interesting problem for future work.

Also of future interest is a more in-depth investigation on the effects of including the model variance as a covariate in the radial variance function. Since v is treated as a function of x, there may be an issue with identifiability of parameters between with the $\log(x)$ term. Better understanding of this relationship and the potential effect it has on estimation of the model variance function is needed.

CHAPTER 4

VARIATIONAL APPROXIMATION FOR SAXS DATA

We now shift our attention to the goal of providing fast estimation procedures for the models described in previous sections. As with many other complex hierarchical models, traditional MCMC-type methodologies suffer from requiring significant computational resources in order to assure posterior convergence. This is particularly true for complicated hierarchical models such as those described in Sections 3.2 and 3.3. Often these models require both significant run-times and substantial user-specified tuning in order to reach acceptable convergence. This chapter highlights the use of variational approximations, a class of deterministic approximation methods, as a means of providing fast, approximate inference for the models described in Chapter 3.

Variational approximations allow for the approximate computation of posterior parameter distributions of complex models in a deterministic fashion (Wainwright and Jordan, 2008). The term *mean field variational Bayes* (MFVB) refers to a class of variational approximations to posterior distributions under a nonparametric product density constraint (Ormerod and Wand, 2010). Applications of MFVB approximations include mixtures (Attias, 2000), linear mixed models (Ormerod and Wand, 2010), and nonparametric regression with missing data (Faes et al., 2011). Given the connection between MFVB and full posterior parameter conditionals, variational approximations can be directly implemented in situations where all parameters have known posterior conditionals and Gibbs sampling is appropriate (Casella and George, 1992). This has led to useful software packages such as `Infer.NET` that allow for "black box" computation of variational approximations given fully known posterior conditionals (Minka et al., 2010; Wang and Wand, 2011).

In this chapter, Section 4.1 provides background on the theory of variational approximations, particularly those performed under product-density restrictions. Section 4.2 describes the variational approximation for traditional mixed models. Section 4.3 introduces a novel variational approximation of the semiparametric regression model with heteroskedastic errors described in Section 2.2. Section 4.5 presents a novel variational method for semiparametric regression via penalized splines with a spatially adaptive penalty term. Finally, Section 4.4 presents work on applying variational approximations to the joint mean-variance model described in Section 3.3.

4.1 Introduction to Variational Approximations

We first provide a brief background on the theory of variational approximations. Let $\mathbf{y}$ be a response vector and $\boldsymbol{\psi}$ be a vector of parameters. Direct evaluation of $p(\mathbf{y})$ in the posterior parameter distribution

$$p(\boldsymbol{\psi} \mid \mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\psi})}{p(\mathbf{y})} \tag{54}$$

may not be analytically or computationally tractable, often requiring high dimensional integration. A standard approach is to use Markov Chain Monte Carlo (MCMC) methods to approximate $p(\boldsymbol{\psi} \mid \mathbf{y})$. However, as models grow in complexity, the computational resources needed to perform MCMC increase dramatically and thus fast approximations become valuable.

Let $q(\boldsymbol{\psi})$ be an arbitrary density function over the parameter space $\boldsymbol{\psi}$. Then

$$\begin{aligned}
\log p(\mathbf{y}) &= \int_{\boldsymbol{\Psi}} q(\boldsymbol{\psi}) \log p(\mathbf{y}) d\boldsymbol{\psi} \\
&= \int_{\boldsymbol{\Psi}} q(\boldsymbol{\psi}) \log \left(\frac{p(\mathbf{y}, \boldsymbol{\psi}) / q(\boldsymbol{\psi})}{p(\boldsymbol{\psi} \mid \mathbf{y}) / q(\boldsymbol{\psi})} \right) d\boldsymbol{\psi} \\
&= \int_{\boldsymbol{\Psi}} q(\boldsymbol{\psi}) \log \left(\frac{p(\mathbf{y}, \boldsymbol{\psi})}{q(\boldsymbol{\psi})} \right) d\boldsymbol{\psi} + \int_{\boldsymbol{\Psi}} q(\boldsymbol{\psi}) \log \left(\frac{q(\boldsymbol{\psi})}{p(\boldsymbol{\psi} \mid \mathbf{y})} \right) d\boldsymbol{\psi}
\end{aligned}$$

$$\geq \int_{\boldsymbol{\Psi}} q(\boldsymbol{\psi}) \log \left(\frac{p(\mathbf{y}, \boldsymbol{\psi})}{q(\boldsymbol{\psi})} \right) d\boldsymbol{\psi}$$

$$=: \log \underline{p}(\mathbf{y}; q), \tag{55}$$

where the inequality follows from Kullback and Leibler (1951), who show the Kullback-Liebler (K-L) divergence $\int q(\boldsymbol{\psi}) \log \left(\frac{q(\boldsymbol{\psi})}{p(\boldsymbol{\psi}|\mathbf{y})} \right) d\boldsymbol{\psi} \geq 0$ with equality if and only if $q(\boldsymbol{\psi}) = p(\boldsymbol{\psi} \mid \mathbf{y})$. The closer the choice of $q(\boldsymbol{\psi})$ is to the true posterior distribution, the smaller the gap between $p(\mathbf{y})$ and $\underline{p}(\mathbf{y}; q)$. Minimizing the gap between these two quantities is equivalent to minimizing the K-L divergence. Variational approximations are useful when analysis of $\underline{p}(\mathbf{y}; q)$ is easier than $p(\mathbf{y})$, which can be achieved by restricting $q(\boldsymbol{\psi})$ in some manner.

4.1.1 Defining $q(\boldsymbol{\psi})$

The most common restriction imposed on the density $q(\boldsymbol{\psi})$, which we use here, is that for some partition of the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_L)$,

$$q(\boldsymbol{\psi}) = \prod_{l=1}^{L} q_l(\boldsymbol{\psi}_l). \tag{56}$$

Variational approximations that impose such a product-type density restriction on $q(\boldsymbol{\psi})$ are referred to as *mean field variational approximations*. When this restriction is used in a Bayesian setting, the term *mean field variational Bayes* or *variational Bayes* for short has become standard. Often there is a somewhat natural partition of the parameter vector given the structure of the model of interest. Both Titterington (2004) and Ormerod and Wand (2010) discuss issues regarding the choice of the partition. Improper partitions (e.g. partitions where there is a high degree of posterior dependence between parameters across partitions) can lead to poor approximations. Alternative forms of variational approximations can be achieved by constraining $q(\boldsymbol{\psi})$ to a class of known parametric density functions but will not be discussed here.

Adopting the notation in Ormerod and Wand (2010), the optimal variational densities, $q_1^*(\boldsymbol{\psi}_1), q_2^*(\boldsymbol{\psi}_2), \ldots, q_L^*(\boldsymbol{\psi}_L)$, that minimize the K-L divergence described in (55) can be found using the relationship

$$q_l^*(\boldsymbol{\psi}_l) \propto \exp\left\{E_{-\boldsymbol{\psi}_l}\left[\log p(\boldsymbol{\psi}_l \mid \cdot)\right]\right\}. \tag{57}$$

Here $p(\boldsymbol{\psi}_l \mid \cdot)$ is the posterior conditional distribution

$$p(\boldsymbol{\psi}_l \mid \mathbf{y}, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{l-1}, \boldsymbol{\psi}_{l+1}, \ldots, \boldsymbol{\psi}_L). \tag{58}$$

The expected value $E_{-\boldsymbol{\psi}_l}$ is the expectation with respect to all q-densities except for $q(\boldsymbol{\psi}_l)$. The dependence of the optimal q-densities on the posterior conditional distributions indicates a close tie to Gibbs sampling (Casella and George, 1992). When conjugate priors are used for $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L$, the posterior conditionals and the variational densities will have the same distributional forms with different parameterizations.

Implementing a variational Bayes approximation under these conditions amounts to determining a set of *variational parameters* that fully define each q-density. Once these variational parameters are determined, the estimates of the model parameters $\psi_1, \ldots, \psi_L$ are taken from the densities $q_1^*, \ldots q_L^*$, usually as the means associated with each density.

The main criticism of variational Bayes methods is the reliance on conjugate forms for the posterior conditional distributions. This is sometimes seen as a limitation when compared to other approximation methods, most notably integrated nested Laplace approximations (Rue et al., 2009) or Expectation-Propagation algorithms (Minka, 2001). Rue et al. (2009) also highlights potential underestimation of parameter variance in variational approximations for latent Gaussian models, depending informativeness of the given data. These are valid

concerns but the simulation studies presented here do not indicate that this a too significant concern for the approximation methods presented in this work.

4.1.2 Example: Random Sample from Normal Distribution

Before we examine the use of variational Bayes for the models described in previous sections, we first present a simple example where we derive the variational approximation for estimating the mean and variance of a normal distribution. This example was presented in Ormerod and Wand (2010) and provides a good illustration of the underlying mechanics of these problems.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ where

$$X_i \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right). \tag{59}$$

Assign priors to the parameters μ and σ^2 of the form

$$\begin{aligned} \mu &\sim \mathcal{N}\left(\mu_\mu, \sigma_\mu^2\right) \\ \sigma^2 &\sim \mathcal{IG}\left(A_{\sigma^2}, B_{\sigma^2}\right). \end{aligned} \tag{60}$$

Here μ_μ, σ_μ^2, A, and B correspond to fixed hyperparameters. Using standard techniques, the posterior conditionals for each parameter are derived to be

$$\begin{aligned} \mu \mid \sigma^2, \mathbf{X} &\sim \mathcal{N}\left(\frac{n\bar{X}/\sigma^2 + \mu_\mu/\sigma_\mu^2}{n/\sigma^2 + 1/\sigma_\mu^2}, \frac{1}{n/\sigma^2 + 1/\sigma_\mu^2}\right) \\ \sigma^2 \mid \mu, \mathbf{X} &\sim \mathcal{IG}\left(A + \frac{n}{2}, B + \frac{1}{2}\left\|\mathbf{X} - \mu \mathbf{1}_N\right\|\right), \end{aligned} \tag{61}$$

where $\bar{X} = \sum_{i=1}^N X_i/N$ and $\mathbf{1}_N$ is a vector of length N with value 1 at each entry.

In order to build a mean field variational approximation for μ and σ^2, we impose the product-density restriction (as described in (56):

$$q(\mu, \sigma^2) = q_\mu(\mu) q_{\sigma^2}(\sigma^2). \tag{62}$$

Using the relationship described in (57), the optimal forms of q_μ and q_{σ^2} are derived from the posterior conditionals as

$$\begin{aligned} q^*_\mu(\mu) &\propto \exp\left[E_{\sigma^2}\left[\log p(\mu \mid \sigma^2, \mathbf{X})\right]\right] \\ q^*_{\sigma^2}(\sigma^2) &\propto \exp\left[E_\mu\left[\log p(\sigma^2 \mid \mu, \mathbf{X})\right]\right]. \end{aligned} \tag{63}$$

Using the posterior conditionals from (60), the optimal q-density for σ^2 becomes

$$q^*_{\sigma^2}(\sigma^2) \propto (\sigma^2)^{-(A+N/2+1)} \exp\left[-\frac{1}{\sigma^2}\left(B + \frac{1}{2} E_\mu\left[\|\mathbf{X} - \mu \mathbf{1}_N\|^2\right]\right)\right]. \tag{64}$$

This functional form leads to the optimal variational distribution

$$\sigma^2 \overset{q^*}{\sim} \mathcal{IG}\left(A + \frac{N}{2}, B_{q(\sigma^2)}\right). \tag{65}$$

Here $B_{q(\sigma^2)}$ is a variational parameter that can be reduced to the following form using the identity described in Ormerod and Wand (2010):

$$\begin{aligned} B_{q(\sigma^2)} &= B + \frac{1}{2} E_\mu\left[\|\mathbf{X} - \mu \mathbf{1}_N\|^2\right] \\ &= B + \frac{1}{2}\left(\left\|\mathbf{X} - \mu_{q(\mu)} \mathbf{1}_N\right\|^2 + N \sigma^2_{q(\mu)}\right). \end{aligned} \tag{66}$$

The variational parameters $\mu_{q(\mu)}$ and $\sigma^2_{q(\mu)}$ refer to the mean and variance of μ under the corresponding variational distribution.

Using similar techniques and the posterior conditional of μ from (60), we derive the variational distribution for μ as

$$\mu \overset{q^*}{\sim} \mathcal{N}\left(\underbrace{\frac{N\bar{X}E_{\sigma^2}\left[1/\sigma^2\right] + \mu_\mu/\sigma_\mu^2}{NE_{\sigma^2}\left[1/\sigma^2\right] + 1/\sigma_\mu^2}}_{\mu_{q(\mu)}}, \underbrace{\frac{1}{NE_{\sigma^2}\left[1/\sigma^2\right] + 1/\sigma_\mu^2}}_{\sigma^2_{q(\mu)}} \right). \tag{67}$$

Here we denote $E_{\sigma^2}\left[1/\sigma^2\right] = \mu_{q(1/\sigma^2)}$. Using the variational distribution of σ^2 described in (65) and the relationship between Gamma and Inverse Gamma random variables

$$\mu_{q(1/\sigma^2)} = \frac{A + N/2}{B_{q(\sigma^2)}}. \tag{68}$$

Using (65) and (67), we now have approximate distributions for our original parameters of interest μ and σ^2, which are controlled by the variational parameters $\mu_{q(\mu)}$, $\sigma^2_{q(\mu)}$, and $B_{q(\sigma^2)}$. As described in Algorithm 1 of Ormerod and Wand (2010), the optimal forms of the mean field variational approximations derived from (57) can be found via a coordinate-ascent type approach. It should be remembered that "optimal" in this case refers to the selection of variational parameters that minimize the K-L divergence described in (55). This is equivalent to maximizing $\log \underline{p}(\mathbf{y}; q)$.

Algorithm 1 Iterative method for determining the optimal variational distributions of μ and σ^2 for the random Normal sample example.

1: **Initialize:** $B_{q(\sigma^2)} > 0$
2: **repeat**
3: $\quad \sigma^2_{q(\mu)} \leftarrow \left(N\left[\frac{A+N/2}{B_{q(\sigma^2)}}\right] + \frac{1}{\sigma_\mu^2} \right)^{-1}$
4: $\quad \mu_{q(\mu)} \leftarrow \left(N\bar{X}\left[\frac{A+N/2}{B_{q(\sigma^2)}}\right] + \frac{\mu_\mu}{\sigma_\mu^2} \right) \sigma^2_{q(\mu)}$
5: $\quad B_{q(\sigma^2)} \leftarrow B + \frac{1}{2}\left(\left\| \mathbf{x} - \mu_{q(\mu)}\mathbf{1}_N \right\|^2 + N\sigma^2_{q(\mu)} \right)$
6: **until** Convergence is reached. Convergence is assessed by monitoring successive values of $\log \underline{p}(\mathbf{y}; q)$ until the change is negligible.
7: Construct parameter estimates of μ and σ^2 using means of optimal variational distributions.

Algorithm 1 describes an iterative method for obtaining values of the variational parameters $\mu_{q(\mu)}$, $\sigma^2_{q(\mu)}$, and $B_{q(\sigma^2)}$ that minimize the K-L divergence. Convergence of this algorithm is assessed by monitoring the iterative values of $\log \underline{p}(\mathbf{y}; q)$. When the step-wise change in $\log \underline{p}(\mathbf{y}; q)$ becomes negligible (either through absolute or relative terms), μ and σ^2 are estimated according to the means of the variational distributions using the converged variational parameters. Using (55), $\log \underline{p}(\mathbf{y}; q)$ is defined as

$$\log \underline{p}(\mathbf{y}; q) = \frac{1}{2} - \frac{N}{2}\log(2\pi) + \frac{1}{2}\log\left(\frac{\sigma^2_{q(\mu)}}{\sigma^2_\mu}\right) - \frac{(\mu_\mu - \mu_{q(\mu)})^2 + \sigma^2_{q(\mu)}}{2\sigma^2_{q(\mu)}}. \tag{69}$$

To illustrate this example, a simulated dataset consisting of $N = 100$ draws from a $\mathcal{N}(10, 4)$ distribution was constructed. Figure 4.1 compares the resulting variational approximations of μ and σ^2 with histograms of the posterior parameter chains from iteratively sampling from the posterior conditionals via Gibbs sampling. Both methods used hyperparameter values of $\mu_\mu = 0$, $\mu^2_\sigma = 10^4$, $A = 10^{-4}$, and $B = 10^{-4}$. The Gibbs sampler was run for 50000 iterations with an initial burn-in of 5000. Convergence of Algorithm 1 for this example was assessed when $\log \underline{p}(\mathbf{y}; q)$ changed by a value less than 10^{-4} and occurred after three steps. Both methods adequately estimate μ and σ^2 with the variational approximation requiring considerably less computational resources (three variational steps versus 5000 posterior draws).

4.2 Variational Approximation for Mixed Models

We now move on to discuss the use of variational approximations in semiparametric regression models. Recall the standard mixed model form

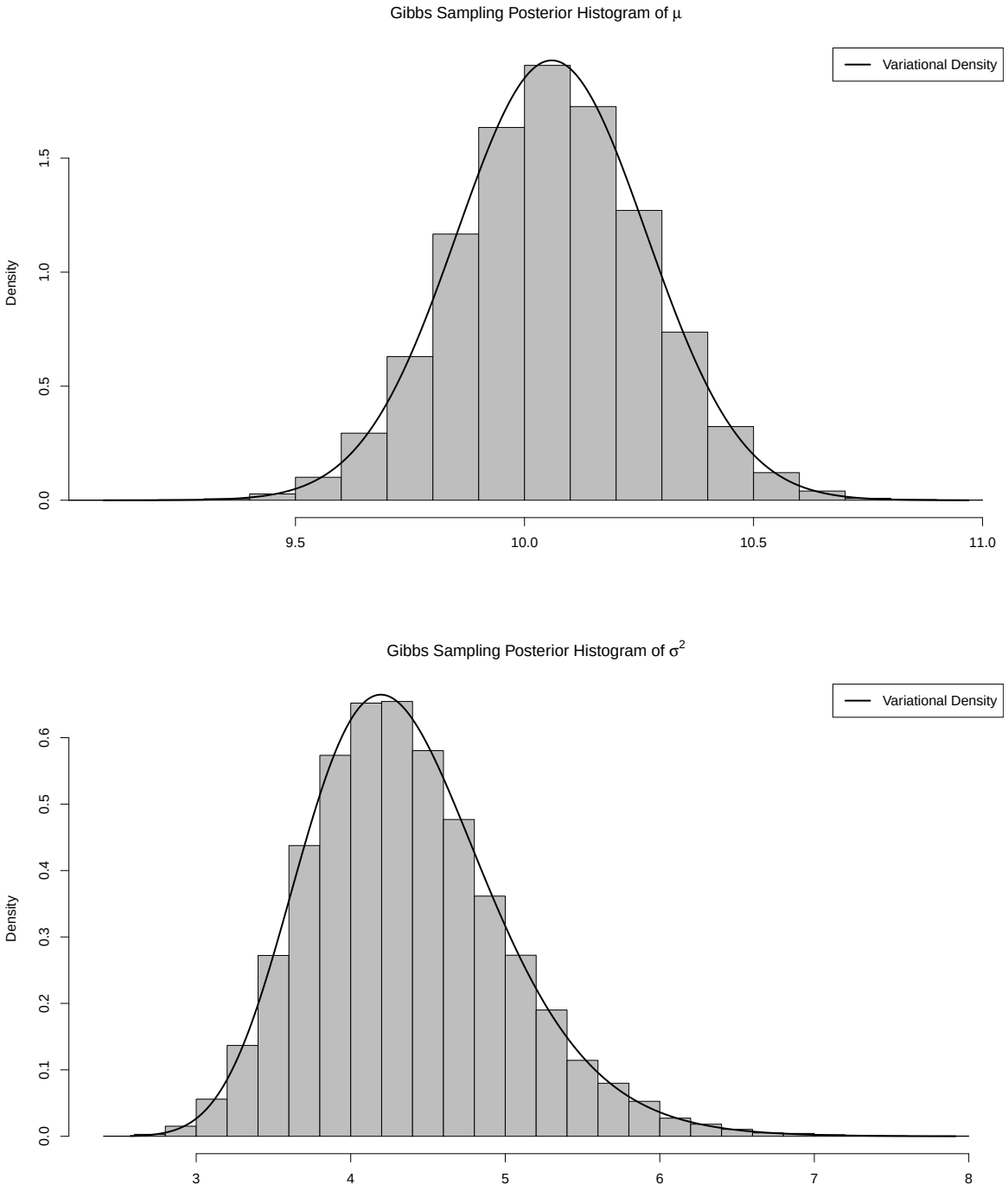


Figure 4.1: Histograms of the Gibbs sampler draws from the posterior conditionals of μ and σ^2 using a run length of 50000 with a burn-in of 5000. The solid black line represents the associated approximate variational density achieved using Algorithm 1.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{b}_1 + \mathbf{Z}_2\mathbf{b}_2 + \cdots + \mathbf{Z}_L\mathbf{b}_L + \epsilon. \tag{70}$$

Here $\mathbf{X}$, $\mathbf{Z}_1$, ..., $\mathbf{Z}_L$ are design matrices with dimensions $N \times p+1$, $N \times K_1$, ..., $N \times K_L$ respectively where N is the number of observations of $\mathbf{Y}$. Define $\boldsymbol{\beta}$ as the vector of all fixed effects. The random effect vectors $\mathbf{b}_1, \ldots, \mathbf{b}_L$ are each considered to follow multivariate Gaussian distributions of the form $\mathcal{N}\left(\mathbf{0}, \sigma^2_{b_m}\mathcal{I}\right)$. The error term ϵ is distributed $\mathcal{N}\left(\mathbf{0}, \sigma^2_\epsilon\mathcal{I}\right)$. As discussed in Section 2.1, there exists a useful link between semiparametric regression and mixed models with Ruppert et al. (2003) providing the standard background on the topic. Given the well-studied nature of mixed models, there exist both likelihood-based and Bayesian frameworks for constructing parameter estimates. For now we will look at a Bayesian version of the linear mixed model in order to construct an appropriate variational approximation. To construct our variational approximation, we use the conjugate priors described in (21) which yield the conditional posterior distributions described in (23) (see Section 3.1 for details).

As with the example shown in Section 4.1.2, the review paper by Ormerod and Wand (2010) leverages the full conjugate structure of the mixed model to produce a simple variational approximation based upon a mean-field assumption. Let $\psi = \left(\boldsymbol{\theta}, \sigma^2_{b_1}, \ldots, \sigma^2_{b_L}, \sigma^2_\epsilon\right)^T$. Assume the variational density $q(\psi)$ has the following product structure:

$$q(\psi) = q_1(\boldsymbol{\theta})q_2(\sigma^2_{b_1}, \ldots, \sigma^2_{b_L}, \sigma^2). \tag{71}$$

Using the relationship described in (57) and the posterior conditionals in (23), the optimal variational densities q_1^* and q_2^* are relatively simple to calculate and take the form

$$\boldsymbol{\theta} \overset{q^*}{\sim} \mathcal{N}\left(\mu_{q(\theta)}, \boldsymbol{\Sigma}_{q(\theta)}\right)$$

$$\begin{aligned}\left(\sigma^2_{b_1}, \ldots, \sigma^2_{b_L}, \sigma^2\right)^T &\overset{q^*}{\sim} \text{Product of Inverse Gamma densities} \\ &= \mathcal{IG}\left(A_\epsilon + \frac{N}{2}, B_{q(\sigma^2_\epsilon)}\right) \times \mathcal{IG}\left(A_1 + \frac{K_1}{2}, B_{q(\sigma^2_{b_1})}\right) \\ &\times \cdots \times \mathcal{IG}\left(A_L + \frac{K_L}{2}, B_{q(\sigma^2_{b_L})}\right)\end{aligned} \tag{72}$$

As with the previous example, these variational densities are defined by a set of variational parameters

$$\begin{aligned}\boldsymbol{\Sigma}_{q(\theta)} &= \left(\frac{A_\epsilon + N/2}{B_{q(\sigma^2_\epsilon)}}\mathbf{C}^T\mathbf{C} + \text{blockdiag}\left(\frac{1}{\sigma^2_\beta}\mathcal{I}_{p+1}, \frac{A_1 + K_1/2}{B_{q(\sigma^2_{b_1})}}\mathcal{I}_{K_1}, \cdots, \frac{A_L + K_L/2}{B_{q(\sigma^2_{b_L})}}\mathcal{I}_{K_L}\right)\right)^{-1} \\ \mu_{q(\theta)} &= \frac{A_\epsilon + N/2}{B_{q(\sigma^2_\epsilon)}}\boldsymbol{\Sigma}_{q(\theta)}\mathbf{C}^T\mathbf{Y} \\ B_{q(\sigma^2_{b_l})} &= B_l + \frac{1}{2}\left(\|\mu_{q(\mathbf{b}_l)}\|^2 + \text{trace}\left(\boldsymbol{\Sigma}_{q(\mathbf{b}_l)}\right)\right) \quad l = 1, \ldots, L \\ B_{q(\sigma^2_\epsilon)} &= B_\epsilon + \frac{1}{2}\left(\|\mathbf{Y} - \mathbf{C}\boldsymbol{\theta}\|^2 + \text{trace}\left(\mathbf{C}^T\mathbf{C}\boldsymbol{\Sigma}_{q(\theta)}\right)\right).\end{aligned} \tag{73}$$

Here the notation $\mu_{q(\mathbf{b}_l)}$ and $\boldsymbol{\Sigma}_{q(\mathbf{b}_l)}$ refers to the portions of the variational mean and covariance matrix of $\boldsymbol{\theta}$ associated with $\mathbf{b}_l$. Algorithm 2 describes the iterative procedure for obtaining variational approximations for the standard mixed model. We omit the derivation of $\log \underline{p}(\mathbf{y}; q)$ here since it is done explicitly in Ormerod and Wand (2010). A generalized `R` script implementing this algorithm for mixed models of any random effect structure was developed and is available from the author upon request.

To illustrate Algorithm 2, consider 100 data points simulated from a true mean function $m(x) = 10\left(\sin\left(\frac{\pi}{2}x\right) + \frac{x^2}{16} - \frac{x^3}{1000}\right)$ with variance $\sigma^2 = 9$ over the region $[0, 2\pi]$. A smooth function

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sum_{k=1}^{10} b_k \left(x_i - \kappa_k\right)_+^2, \tag{74}$$

Algorithm 2 Iterative method for determining the optimal variational distributions for the parameters of a mixed model.

Initialize: $B_{q(\sigma_\epsilon^2)}, B_{q(\sigma^2_{\mathbf{b}_1})}, \ldots, B_{q(\sigma^2_{\mathbf{b}_L})} > 0$

repeat

  $\boldsymbol{\Sigma}_{q(\theta)} \leftarrow \left(\frac{A_\epsilon + N/2}{B_{q(\sigma_\epsilon^2)}} \mathbf{C}^T\mathbf{C} + \text{blockdiag}\left(\frac{1}{\sigma_\beta^2}\mathcal{I}_{p+1}, \frac{A_1+K_1/2}{B_{q(\sigma^2_{b_1})}}\mathcal{I}_{K_1}, \cdots, \frac{A_L+K_L/2}{B_{q(\sigma^2_{b_L})}}\mathcal{I}_{K_L} \right) \right)^{-1}$

  $\mu_{q(\theta)} \leftarrow \frac{A_\epsilon + N/2}{B_{q(\sigma_\epsilon^2)}} \boldsymbol{\Sigma}_{q(\theta)} \mathbf{C}^T\mathbf{Y}$

  for all $l = 1, \ldots, L$ **do**

    $B_{q(\sigma^2_{b_l})} \leftarrow B_l + \frac{1}{2}\left(\|\mu_{q(\mathbf{b}_l)}\|^2 + \text{trace}\left(\boldsymbol{\Sigma}_{q(\mathbf{b}_l)}\right)\right)$

  end for

  $B_{q(\sigma_\epsilon^2)} \leftarrow B_\epsilon + \frac{1}{2}\left(\|\mathbf{Y} - \mathbf{C}\boldsymbol{\theta}\|^2 + \text{trace}\left(\mathbf{C}^T\mathbf{C}\boldsymbol{\Sigma}_{q(\theta)}\right)\right)$

until Convergence is reached. Convergence is assessed by monitoring successive values of $\log \underline{p}(\mathbf{y}; q)$ until the change is negligible.

Construct parameter estimates of $\boldsymbol{\theta}$, σ_ϵ^2, $\sigma_{b_1}^2, \ldots, \sigma_{b_L}^2$ using means of converged variational distributions.

was built from a set of truncated quadratic basis functions, with 10 knots $\{\kappa_k\}_{k=1}^1 0$ corresponding equally-spaced quantiles of $\{x_i\}_{i=1}^N$ ranging from 0.1 to 0.9. Convergence for this variational approximation was achieved in six steps with a run time of approximately 0.09 seconds. Figure 4.3 compares the variational approximation to the fit from a Gibbs sampler that directly draws from the posterior conditional forms found in (23). The runtime for 1000 draws using this Gibbs sampler was approximately 104 seconds. In addition to the Bayesian formulation, a fit using standard frequentist mixed model software, specifically the *lme* function from the *nlme* package in *R*, is also compared. For a standard mixed model, the *lme* function's run time was comparable to the variational approximation. The fit associated with the variational approximation compares quite well to both the alternative methods, which are all nearly graphically indistinguishable when plotted together.

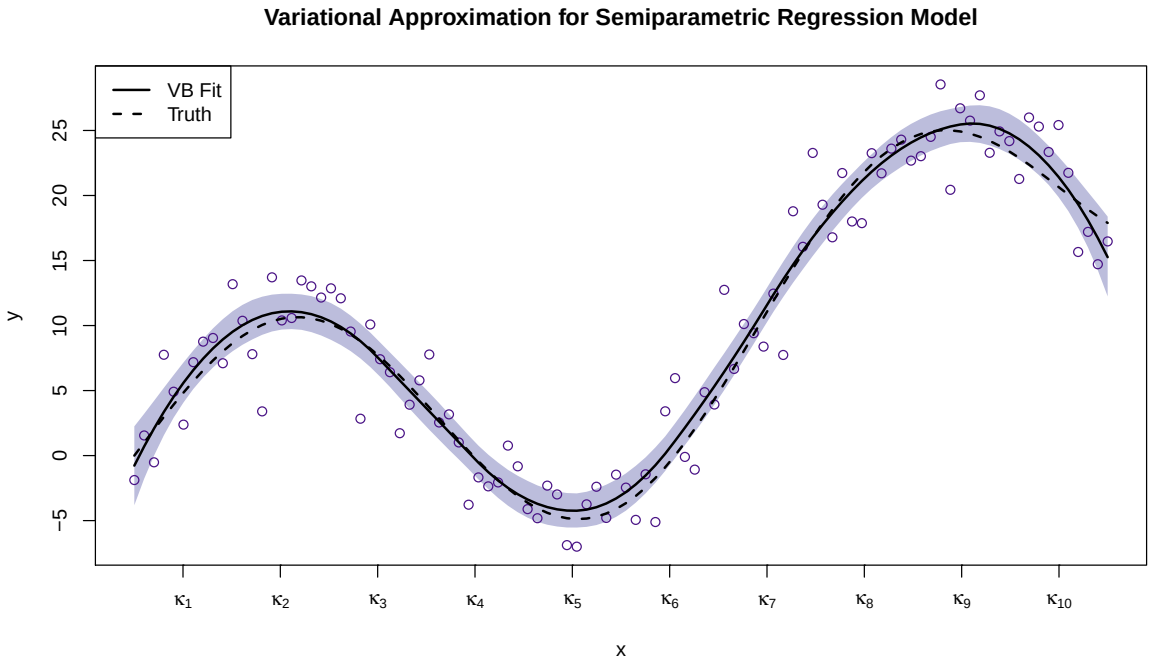


Figure 4.2: This figure shows the result of Algorithm 2 applied to simulated data with true mean function $m(x) = 10\left(\sin\left(\frac{\pi}{2}x\right) + \frac{x^2}{16} - \frac{x^3}{1000}\right)$ and variance $\sigma^2 = 9$ for $x \in [0, 2\pi]$. Knots $\kappa_1, \ldots, \kappa_{10}$, placed at equally spaced quantiles for $\{x_i\}_{i=1}^N$ ranging from 0.1 to 0.9, are used for a truncated quadratic basis. The shaded region represents a point-wise 95% credible region calculated using the variational distribution of $\boldsymbol{\theta}$.

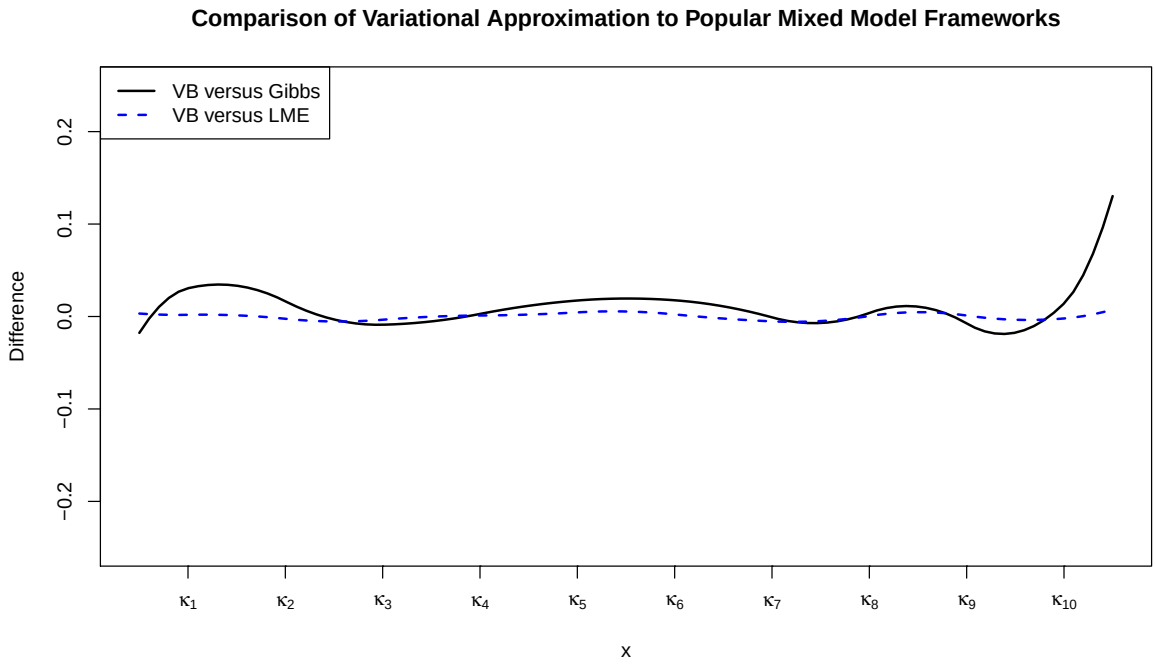


Figure 4.3: The black solid line represents the difference in fit between the variational approximation in Figure 4.2 and a Gibbs sampler implementation run for 1000 iterations with a burn-in of 100. The blue dashed line represents the difference between the variational approximation and a standard frequentist fit calculated using the *lme* function from the *nlme* package in *R*. Both differences are very close to 0 over all values of x.

4.3 Laplace Variational Approximation for Semiparametric Regression with Heteroskedastic Errors

We now approximation method for semiparametric regression in the presence of heteroskedastic errors, described in (13), via variational Bayes. Nott et al. (2012) describe a variational approximation for the standard linear model in the presence of heteroskedasticity of known parametric form. Our approximation applies more broadly, to linear mixed models generally and to semiparametric regression under heteroskedasticity specifically. Unlike Nott et al. (2012), we do not assume a parametric form for the heteroskedasticity, but model it flexibly with penalized splines.

Let $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\theta}_V, \sigma^2_{b_1}, \ldots, \sigma^2_{b_L}, \sigma^2_{c_1}, \ldots, \sigma^2_{c_M})^T$ be a vector containing all parameters of interest. Recall that $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{b}_1, \ldots, \mathbf{b}_L)^T$ and $\boldsymbol{\theta}_V = (\boldsymbol{\delta}, \mathbf{c}_1, \ldots, \mathbf{c}_M)^T$ are vectors of model coefficients for the mean and variance levels respectively. Using the product density constraint from (56), we assume that the variational density $q(\boldsymbol{\psi})$ is of the form

$$q(\boldsymbol{\psi}) = q_1(\boldsymbol{\theta}) q_2(\boldsymbol{\theta}_V) q_3(\sigma^2_{b_1}, \ldots, \sigma^2_{b_L}) q_4(\sigma^2_{c_1}, \ldots, \sigma^2_{c_M}). \tag{75}$$

For notational simplicity, the subscript index on each component density of $q(\boldsymbol{\psi})$ is dropped in future references (e.g. $q_1(\boldsymbol{\theta}) = q(\boldsymbol{\theta})$). The notation $\overset{q^*}{\sim}$ is used to describe the optimal distributions of a parameter derived using the relationship described in (57).

Under this assumption, the relationship described in (57) yields an optimal variational density q^* of the form

$$\begin{aligned}
\theta &\overset{q^*}{\sim} \mathcal{N}\left(\mu_{q(\theta)}, \boldsymbol{\Sigma}_{\mathbf{q}(\theta)}\right) \\
q^*(\sigma^2_{b_1}, \ldots, \sigma^2_{b_L}) &= \text{Product of } \mathcal{IG}\left(A_{b_l} + \frac{K_l}{2}, B_{q(\sigma^2_{b_l})}\right) \text{ densities for } l = 1, \ldots, L \\
q^*(\sigma^2_{c_1}, \ldots, \sigma^2_{c_M}) &= \text{Product of } \mathcal{IG}\left(A_{c_l} + \frac{K_{V_m}}{2}, B_{q(\sigma^2_{c_m})}\right) \text{ densities for } m = 1, \ldots, M
\end{aligned}$$

$$q^*(\boldsymbol{\theta}_V) \propto \exp\Bigg(-\frac{1}{2}\Bigg(\sum_{i=1}^{N} \left[(y_i - \mathbf{C}_i^T \mu_{q(\boldsymbol{\theta})})^2 + \mathbf{C}_i^T {\boldsymbol{\Sigma}_{\mathbf{q}(\boldsymbol{\theta})}}^T \mathbf{C}_i\right] \exp\{-\mathbf{C}_{V_i}^T \boldsymbol{\theta}_V\} + \sum_{i=1}^{N} \mathbf{C}_{V_i}^T \boldsymbol{\theta}_V + {\boldsymbol{\theta}_V}^T \boldsymbol{\Omega} \boldsymbol{\theta}_V \Bigg)\Bigg). \tag{76}$$

Let $\mu_{q(1/\sigma^2_{b_l})} = E_{-\boldsymbol{\theta}}\left[1/\sigma^2_{b_l}\right]$ and $\mu_{q(1/\sigma^2_{c_m})} = E_{-\boldsymbol{\theta}_V}\left[1/\sigma^2_{c_m}\right]$. Let

$$\boldsymbol{\Omega} \quad = \text{blockdiag}\left(\sigma_\delta^{-2}\mathcal{I}_r, \sigma_{c_1}^{-2}\mathcal{I}_{K_{V_1}}, \ldots, \sigma_{c_M}^{-2}\mathcal{I}_{K_{V_M}}\right). \tag{77}$$

The variational covariance matrix associated with $\boldsymbol{\theta}$ is

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \quad = \left(\text{blockdiag}(\tfrac{1}{\sigma^2_\beta}\mathcal{I}_p, \mu_{q(1/\sigma^2_{b_1})}\mathcal{I}_{K_1}, \ldots, \mu_{q(1/\sigma^2_{b_L})}\mathcal{I}_{K_L}) + \mathbf{C^T \Gamma C}\right)^{-1}, \tag{78}$$

where $\boldsymbol{\Gamma}$ is defined as the diagonal matrix

$$\boldsymbol{\Gamma} = \text{diag}\Big(E_{-\boldsymbol{\theta}}\left[\exp\left\{-\mathbf{C}_{V_1}^T\boldsymbol{\theta}_V\right\}\right], \ldots, E_{-\boldsymbol{\theta}}\left[\exp\left\{-\mathbf{C}_{V_N}^T\boldsymbol{\theta}_V\right\}\right]\Big). \tag{79}$$

The variational mean parameter associated with $q(\boldsymbol{\theta})$ is

$$\mu_{q(\boldsymbol{\theta})} = \boldsymbol{\Sigma}_{\mathbf{q}(\boldsymbol{\theta})}\mathbf{C^T \Gamma y}. \tag{80}$$

It is important to note that the diagonal elements of the matrix $\boldsymbol{\Gamma}$,

$$\gamma_i = E_{-\boldsymbol{\theta}}\left[\exp\left\{-\mathbf{C}_{V_i}^T\boldsymbol{\theta}_V\right\}\right], \tag{81}$$

correspond to the moment generating function of $\boldsymbol{\theta}_V$ under the distribution described by $q(\boldsymbol{\theta}_V)$ evaluated at the row vector $-\mathbf{C}_{V_i}^T$. This relationship is important since it gives us a closed form evaluation of this expected value in terms of the variational parameters of $q(\boldsymbol{\theta}_V)$ for a large class of variational densities.

For each $l = 1, \ldots, L$ and $m = 1, \ldots, M$, $B_{q(\sigma^2_{b_l})} = B_{b_l} + \|\mu_{q(b_l)}\|^2/2 + \text{trace}(\Sigma_{q(b_l)})/2$ and $B_{q(\sigma^2_{c_m})} = B_{c_m} + \|\mu_{q(c_m)}\|^2/2 + \text{trace}(\Sigma_{q(c_m)})/2$ respectively. The notation $\mu_{q(b_l)}$ and $\mu_{q(c_m)}$ refer to the components the variational means $\mu_{q(\boldsymbol{\theta})}$ and $\mu_{q(\boldsymbol{\theta}_V)}$ associated with $\mathbf{b_l}$ and $\mathbf{c_m}$. Similar notation applies to the variational covariance matrices $\Sigma_{q(b_l)}$ and $\Sigma_{q(c_m)}$. Appendix A contains a detailed derivation of the variational density q for the model described in (13).

The lack of conjugate structure for describing the parameters associated with the smooth fit of the $\log(\sigma^2)$ function, $\boldsymbol{\theta}_V$, that was a source of difficulty for traditional approaches for estimating the heteroskedastic semiparametric regression model described in (13), causes issues for the derivation of a variational approximation as well. The variational distribution of $q(\boldsymbol{\theta}_V)$ described in (76) does not have a known form. For traditional computational methods, this can be handled in a number of ways. As discussed previously, the methods described in Baladandayuthapani et al. (2005) and Crainiceanu et al. (2007) modify their models by adding a latent noise parameter to the smooth function described by $\boldsymbol{\theta}_V$. This allows for posterior conditional structures that amount to N univariate Metropolis-Hasting steps for updating each σ_i^2. If we attempt to build a variational approximation for this modified model, we are left with N univariate variational densities $q(\sigma_1^2)$, $q(\sigma_2^2)$, $\ldots$, $q(\sigma_N^2)$, each with unknown form. Unknown variational forms for univariate parameters can sometimes be handled through a discretized,"Griddy Gibbs" type method (Pham et al., 2013; Ritter and Tanner, 1992). However, we are concerned about the ability of this method to scale with sample size in this case since N discretizations would need to be performed. Because of this we keep our focus on dealing directly with the multivariate distribution of $\boldsymbol{\theta}_V$.

A discretization-based strategy for dealing with the unknown multivariate variational density $q(\boldsymbol{\theta}_V)$ would be too computationally intensive for a "fast" approximation method, especially since the dimensionality of $\boldsymbol{\theta}_V$ is tied to the complexity of the model for the variance structure and the number of knots being used. Instead, we leverage the fact that $q(\boldsymbol{\theta}_V)$ in

(76) has a Gaussian structure save for the term

$$-\frac{1}{2}\sum_{i=1}^{N}\left[(y_i-\mathbf{C}_i^T\mu_{q(\boldsymbol{\theta})})^2+\mathbf{C}_i^T\boldsymbol{\Sigma}_{\mathbf{q}(\boldsymbol{\theta})}\mathbf{C}_i\right]\exp\left(-\mathbf{C}_{V_i}^T\boldsymbol{\theta}_V\right). \tag{82}$$

We propose to use an approximate form of $q^*(\boldsymbol{\theta}_V)$, denoted $\tilde{q}^*(\boldsymbol{\theta}_V)$, derived via Laplace approximation to construct a variational approximation for the model described in (13).

4.3.1 Laplace Approximation for $\boldsymbol{\theta}_V$

As in Nott et al. (2012) and Wang and Blei (2012), we consider a Laplace approximation to replace $q(\boldsymbol{\theta}_V)$ with a multivariate Gaussian density. First, let $q^*(\boldsymbol{\theta}_V)=\exp\left[-h(\boldsymbol{\theta}_V\mid\cdot)\right]$ and consider the Taylor approximation of the function $h(\boldsymbol{\theta}_V\mid\cdot)$ about some value α:

$$h(\boldsymbol{\theta}_V\mid\cdot)\approx h(\alpha\mid\cdot)+(\boldsymbol{\theta}_V-\alpha)^T\mathbf{J}_{\boldsymbol{\theta}_\mathbf{V}}(\alpha)+\frac{1}{2}(\boldsymbol{\theta}_V-\alpha)^T\mathbf{H}_{\boldsymbol{\theta}_\mathbf{V}}(\alpha)(\boldsymbol{\theta}_V-\alpha). \tag{83}$$

Here $\mathbf{J}_{\boldsymbol{\theta}_\mathbf{V}}(\alpha)$ corresponds to the Jacobian vector of first derivatives of h with respect to $\boldsymbol{\theta}_V$ evaluated at α and $\mathbf{H}_{\boldsymbol{\theta}_\mathbf{V}}(\alpha)$ corresponds to the Hessian matrix containing all partial second derivatives of h evaluated at α. If $\alpha=\text{argmin}\, h(\boldsymbol{\theta}_V\mid\cdot)$, then the approximate form of $q^*(\boldsymbol{\theta}_V)$, which we denote $\tilde{q}^*(\boldsymbol{\theta}_V)$, becomes

$$\tilde{q}^*(\boldsymbol{\theta}_V)\propto\exp\left[-\frac{1}{2}(\boldsymbol{\theta}_V-\alpha)^T\frac{1}{2}\mathbf{H}_{\boldsymbol{\theta}_\mathbf{V}}(\alpha)(\boldsymbol{\theta}_V-\alpha)\right]. \tag{84}$$

This implies that $\boldsymbol{\theta}_V$ has an approximate variational distribution of the form

$$\boldsymbol{\theta}_V\overset{\tilde{q}^*}{\sim}\mathcal{N}\left(\alpha,\left(\frac{1}{2}\mathbf{H}_{\boldsymbol{\theta}_\mathbf{V}}(\alpha)\right)^{-1}\right). \tag{85}$$

Using this Gaussian form, the diagonal elements of Γ become

$$\gamma_i=\exp\left(-\mathbf{C}_{V_i}^T\mu_{q(\boldsymbol{\theta}_V)}+.5\mathbf{C}_{V_i}^T\Sigma_{q(\boldsymbol{\theta}_V)}\mathbf{C}_{V_i}^T\right), \tag{86}$$

and yields fully expressed formulas for $\mu_{q(\theta)}$ and $\Sigma_{q(\theta)}$.

4.3.2 Algorithm

Using the q-densities described in (76) and (85), constructing a variational Bayes approximation of our model consists of updating the parameters associated with each q-density until $q(\psi)$ has sufficiently approximated $p(\psi \mid \cdot)$. As before, this is done via a coordinate ascent type algorithm (Ormerod and Wand, 2010; Nott et al., 2012; Wang and Blei, 2012). The variational parameters of interest here are denoted as $\mu_{q(\boldsymbol{\theta})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$, $\mu_{q(\boldsymbol{\theta}_V)}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta}_V)}$, $B_{q(\sigma^2_{b_1})}$, $\ldots$, $B_{q(\sigma^2_{b_L})}$, $B_{q(\sigma^2_{c_1})}$, $\ldots$, and $B_{q(\sigma^2_{c_M})}$. During each step of the algorithm, the most recent parameters are used and convergence is not assessed until each full cycle has been completed. Given the Inverse Gamma variational distribution of $\sigma^2_{b_l}$ and $\sigma^2_{c_m}$, $\mu_{q(1/\sigma^2_{b_l})} = (A_{b_l} + K_l/2)\, B_q(\sigma^2_{b_l})^{-1}$ and $\mu_{q(1/\sigma^2_{c_m})} = (A_{c_m} + K_{V_m}/2)\, B_q(\sigma^2_{c_m})^{-1}$ for all $l = 1, \ldots, L$ and $m = 1, \ldots M$.

Algorithm 3 describes the process of obtaining the variational parameter estimates associated with the approximation described in (76).

Similarly to the other algorithms presented here, the updating steps in Algorithm 3 are completely deterministic. Each iteration through the algorithm requires the determination of at least a local minimum of $h(\boldsymbol{\theta}_V \mid \cdot)$, α. This is currently done through the application of off-the-shelf numerical optimization algorithms, namely the *optim* command in *R*. This multidimensional minimization problem does cause a slight slow-down for this variational approximation compared to problems where an approximation for a fully-conjugate model is being constructed. However, practical implementations of Algorithm 3 show that this is only of mild concern. This is largely due to the fact that variational approximations tend to converge quickly and thus only a small number of optimization steps are needed.

Algorithm 3 Iterative method for determining the optimal variational distributions for the parameters of a heteroskedastic semiparametric regression via penalized splines model

Initialize: $\mu_{q(\boldsymbol{\theta})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$, $B_{q(\sigma^2_{b_1})}$,..., $B_{q(\sigma^2_{b_L})}$,$B_{q(\sigma^2_{c_1})}$,..., $B_{q(\sigma^2_{c_M})}$

repeat

  Compute $\alpha = \text{argmin } h(\boldsymbol{\theta}_V \mid \cdot)$, $\mathbf{H}_{\tilde{q}(\boldsymbol{\theta}_V)}(\alpha)$

  $\mu_{\tilde{q}(\boldsymbol{\theta}_V)} \leftarrow \alpha$

  $\boldsymbol{\Sigma}_{\tilde{q}(\boldsymbol{\theta}_V)} \leftarrow \left(\frac{\mathbf{H}_{\boldsymbol{\theta}_V}(\alpha)}{2}\right)^{-1}$

  for all $m = 1, \ldots, M$ **do**

    $B_{q(\sigma^2_{c_m})} \leftarrow B_{c_m} + \frac{1}{2}\|\mu_{q(\mathbf{c_m})}\|^2 + \frac{1}{2}\text{trace}(\boldsymbol{\Sigma}_{q(\mathbf{c_m})})$

  end for

  $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \leftarrow \left(\text{blockdiag}(\frac{1}{\sigma^2_\beta}\mathcal{I}_p, \mu_{q(1/\sigma^2_b)}\mathcal{I}_K) + \mathbf{C^T\Gamma C}\right)^{-1}$

  $\mu_{q(\boldsymbol{\theta})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}C^T\Gamma y$

  for all $l = 1, \ldots, L$ **do**

    $B_{q(\sigma^2_{b_l})} \leftarrow B_{b_l} + \frac{1}{2}\|\mu_{q(\mathbf{b_l})}\|^2 + \frac{1}{2}\text{trace}(\boldsymbol{\Sigma}_{q(\mathbf{b_l})})$

  end for

until Convergence is Reached (See Section 4.3.3)

Construct parameter estimates using mean of variational distributions

4.3.3 Convergence Monitoring

In fully conjugate variational models such as those in Ormerod and Wand (2010), coordinate ascent algorithms are guaranteed to converge to at least locally optimal values of the variational parameters. Use of an approximate q-density for $\boldsymbol{\theta}_V$ removes this guarantee. Since the only assumption being made is that of a quadratic approximation of $h(\boldsymbol{\theta}_V \mid \cdot)$ and $h(\boldsymbol{\theta}_V \mid \cdot)$ has the form of quadratic plus an extra term, we conjecture that parameter updates from Algorithm 3 will converge; empirical results described below support this conjecture. A common objective function to monitor for convergence is $\log \underline{p}(\mathbf{y}; q)$ defined in (55). Wang and Blei (2012) suggest monitoring $\|\mu_{q(\boldsymbol{\theta}_V)}\|$ to assess convergence as well. Appendix B contains the derivation of $\log \underline{p}(\mathbf{y}; q)$. A more in-depth look at the effect of the Gaussian approximation for $q(\boldsymbol{\theta}_V)$ on the convergence properties of $\log \underline{p}(\mathbf{y}; q)$ is a topic of future work.

4.3.4 Covariance Matrices

$10,000$ iterations The matrices to be inverted in steps 5 and 7 of Algorithm 3 may not be positive definite, perhaps due to poor initial conditions. If an invalid covariance matrix is proposed in a step, we use a generalized inverse and a ridge adjustment, adding to each diagonal entry of the matrix to be inverted the absolute value of its smallest eigenvalue. Usually this is not needed for most application of the variational approximation to the heteroskedastic semiparametric regression model. Investigating the effect of these matrix structures on the underlying uncertainty estimates is an open question. Also of interest would be potentially including some constraints that ensured the estimated matrices were both symmetric and positive definite, although the implementation of this is not directly obvious.

4.3.5 Single Curve Simulation Examples

To illustrate our methodology, we first consider two simulated datasets of $N = 200$ both with a true mean function of $m(x) = -\frac{1}{8}(x-5)^3 + x$ and with true variance functions of $v_1(x) = \left(\frac{1}{4}x + \frac{1}{2}\right)^3$ or $v_2(x) = \exp\left\{\frac{(x-5)^2}{5}\right\}$, with $x \in [0, 10]$. We fit a model of the form described in (13) using a piecewise quadratic basis ($p = r = 2$). For both the mean and variance levels, $K = K_V = 10$ knots were placed at equally spaced quantiles of x (Ruppert et al., 2003). The hyperparameters associated with the priors for σ_b^2 and σ_c^2 are $A_b = B_b = A_c = B_c = 10^{-5}$. The variance hyperparameters associated with the fixed effects β and δ are $\sigma_\beta^2 = \sigma_\delta^2 = 10^5$.

Before the algorithm is started, an initial fitting of $\mathbf{y}$ under a homoskedastic assumption is performed using the same model structure. The corresponding $\widehat{\boldsymbol{\theta}}$ is used as an initial value of $\mu_{q(\boldsymbol{\theta})}$. The matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$ is initialized using a diagonal matrix of appropriate scale for the data, and $B_q(\sigma_b^2)$ and $B_q(\sigma_c^2)$ are initialized to 1/100.

Figure 4.4 shows the resulting fit of the simulated data under both variance structures using our variational approximation. The approximation closely follows the true mean function in both examples but encounters increased difficulty in areas with higher noise. The shaded areas represent 95% pointwise credible bounds determined by taking a sample of size 10000 from the converged variational distribution of $q^*(\boldsymbol{\theta})$. This measure of uncertainty reflects the non-constant variance structure with wider bands being associated with areas of higher variance. The true mean function $m(x)$ was fully covered by the pointwise credible bounds associated with each variance function.

Figure 4.5 shows the estimated log variance functions $\log(v_1(x))$ and $\log(v_2(x))$ associated with the data simulated from $m(x)$. The estimated curves associated with our variational approximation (solid lines) closely follow the the true log variance functions (dashed lines).

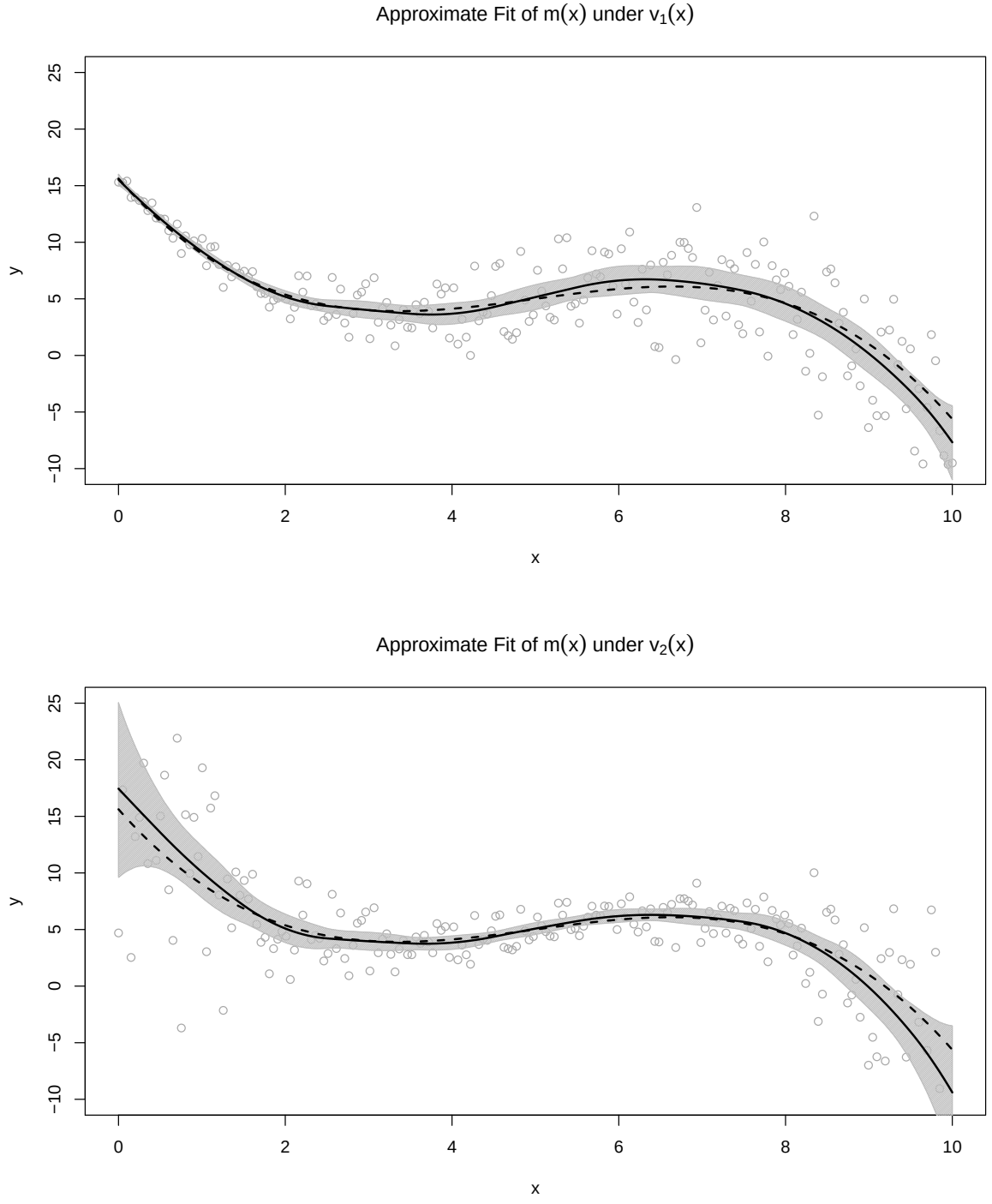


Figure 4.4: Variational approximation of simulated data with true mean function $m(x) = -\frac{1}{8}(x-5)^3 + x$ (solid line) and variance functions $v_1(x) = \left(\frac{1}{4}x + \frac{1}{2}\right)^3$ (top) and $v_2(x) = \exp\left\{\frac{(x-5)^2}{5}\right\}$ (bottom). The shaded area corresponds to the 95% pointwise credible bounds calculated from 10000 draws from the variational distribution $q^*(\boldsymbol{\theta})$. The true mean function is represented by a dashed line in each plot.

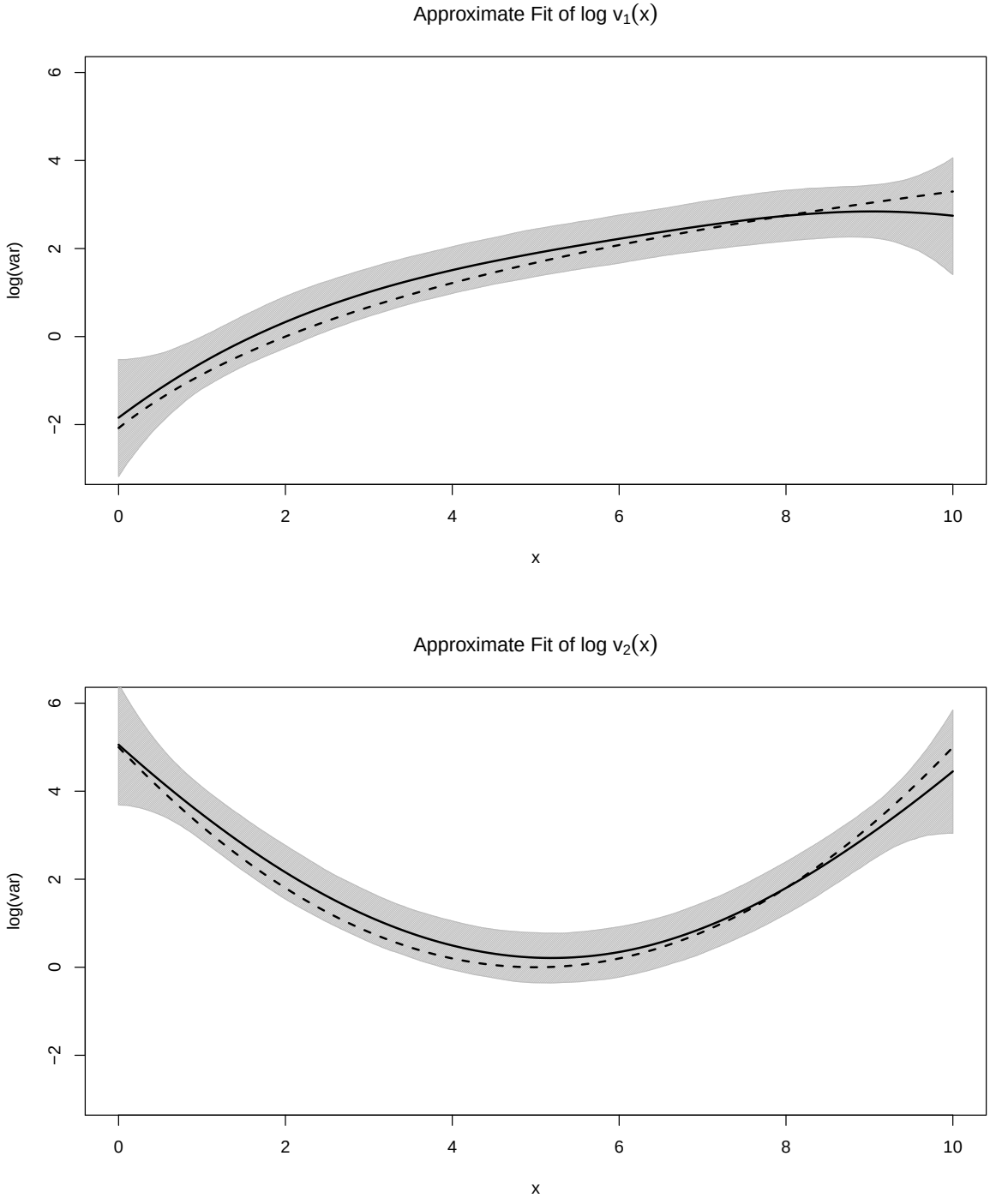


Figure 4.5: Variational approximation fits (solid line) of $\log v_1(x)$ (top) and $\log v_2(x)$ (bottom) associated with simulated data from the true mean function $m(x) = -\frac{1}{8}(x-5)^3 + x$. The shaded area corresponds to the 95% pointwise credible bounds calculated from 10000 draws from the variational distribution $q^*(\boldsymbol{\theta}_V)$. The true variance functions are represented by dashed lines.

The 95% pointwise credible bounds calculated using 10000 draws from the variational distribution $q^*(\boldsymbol{\theta}_V)$ cover the true variance function over the range of x in both cases.

4.3.6 Single Curve Simulation Comparison to MCMC Methods

We are using an approximate form of the optimal variational density associated with $\boldsymbol{\theta}_V$ and this variational method is itself an approximation. We now consider the quality of our approximation, as compared to the posterior distribution evaluated using a combination of Gibbs sampling and an appropriate multivariate posterior sampling algorithm, such as DRAM (see Section 3.2). Figure 4.6 compares the 95% pointwise credible intervals for both the mean and log variance curves using a simulated sample of 1000 points with true curves $m(x)$ and $v_1(x)$. The MCMC procedure was run for $10,000$ iterations with a burn-in length of 10%. The DRAM step consists of two stages of potential proposal acceptance at each iteration with covariance adaptation every 100 steps. All other model choices (basis functions, knots, etc.) are the same as the previous simulated results.

When comparing the coverage of the true mean curve, both the variational approximation and the MCMC method are very close. This is not surprising given the conjugate structure of the $\boldsymbol{\theta}$ parameter which controls the fit estimate of the $m(x)$. The posterior conditional described in (29) for $\boldsymbol{\theta}$ is multivariate Gaussian as is the variational density $q^*_{\boldsymbol{\theta}}$. The coverage comparison for the log variance curve is more involved. As in Figure 4.5, the credible intervals from the variational approximation behave smoothly and cover the true function $\log v_1(x)$. The credible bounds from the MCMC procedure have a more pronounced curvature and are narrower in some places. Also, the MCMC bands have some trouble covering $\log v_1(x)$, particularly in the region between 0 and 2. This behavior can most likely be attributed to the difficulty of the multivariate sampling problem for the posterior conditional distribution $p(\boldsymbol{\theta}_V \mid \cdot)$. In this case $p(\boldsymbol{\theta}_V \mid \cdot)$ describes a 13-dimensional distribution from which we have to draw samples. While DRAM performs significantly better than standard Metropolis-

Hastings for this problem, it can still have difficulty fully covering the parameter space, especially if the space is complicated and potentially multi-modal. Improvements in fit may be achievable through considerable tuning (modifying adaption length, covariance scaling, etc.) or implementing a more sophisticated algorithm to handle the multivariate sampling problem. These results highlight the additional advantage that variational approximations provide in that they do not require specification of tuning parameters.

4.3.7 Computational Performance of Single Curve Examples

Convergence of Algorithm 3 is assessed by monitoring the relative change of $\log \underline{p}(\mathbf{y}; q)$. Variational algorithms are known to converge very quickly to appropriate estimates (Ormerod and Wand, 2010). Table 4.1 contains the run time for the variational approximation associated with each variance curve at increasing sample sizes. All computations were performed on a MacBook Pro with a 2.3 GHz Intel Core i5 processor and 4 GB of RAM. For comparison, the run time for $10,000$ iterations of the MCMC procedure described previously are included in Table 4.1. Our algorithm is slower than those that operate in fully conjugate situations since we have a numerical optimization embedded at each step to estimate α from the most current $h(\boldsymbol{\theta}_V \mid \cdot)$. However, there is still a dramatic speed increase compared to the MCMC method with run times for the simulated cases being approximately 5 to 17 times faster, depending on sample size, using the variational approximation. This reduction of computational cost coupled with the lack of manual tuning parameters demonstrate the value of these variational approximations. Further computational improvements may be achieved through the use of a custom optimization routine rather than off-the-shelf implementations.

4.3.8 Vertically Shifted Curves with Common Variance Structure

The following sections highlights selected examples that expand on the single curve simulation problems to include more complicated semiparametric structures at both the mean and variance level. As before, all hyperparameters for fixed effect variances are set at

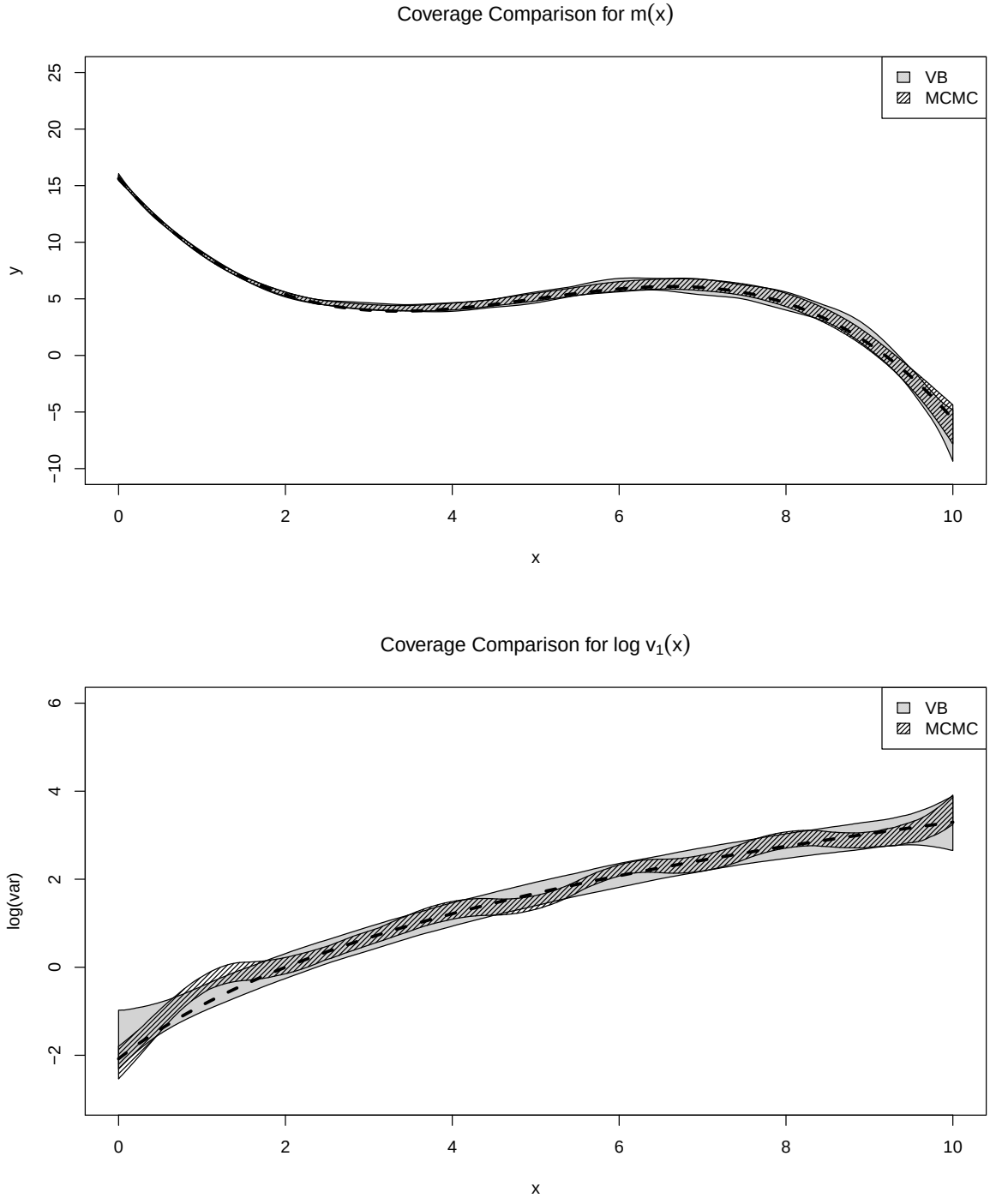


Figure 4.6: The solid light gray regions represent the pointwise 95% credible bounds drawn from the approximating variational distributions $q_{\boldsymbol{\theta}}$ and $\tilde{q}_{\boldsymbol{\theta}_V}$ for a simulated data set of 1000 draws with $m(x)$ and $v_1(x)$ as the true mean and variance functions. The line shaded regions correspond to the resulting pointwise posterior 95% credible bounds taken from the hybrid Gibbs-DRAM MCMC procedure using the posterior conditionals described in (29) run for 10000 iterations with a 10% burn-in length.

Table 4.1: Run times (seconds) for both the variational approximations and MCMC procedures described in Section 4.3.5 for various sample sizes. The MCMC procedure was run for 10,000 iterations in each case.

Method	Sample Size (N)			
	200	400	800	1600
Variational approximation for $v_1(x)$	10.68	20.02	39.35	98.67
Variational approximation for $v_2(x)$	10.73	21.18	42.18	95.30
MCMC for $v_1(x)$	57.89	116.863	315.89	1702.41
MCMC for $v_2(x)$	54.22	124.67	349.96	1386.03

$\sigma_\beta^2 = \sigma_\delta^2 = 10^5$ and the rate and scale (A and B) hyperparameters for the Inverse Gamma priors associated with the random effect variances are set at 10^{-5}. Truncated quadratic bases of the form $(x-\kappa)_+^2$ are used for all nonparametric fits unless otherwise specified. Fifteen knots, denoted $\kappa_1, \ldots, \kappa_{15}$, selected using equally spaced quantiles ranging from 5% to 95% are used to evaluate the truncated quadratic spline basis functions at both the mean and variance levels.

Our first example considers the case where we have two curves whose only apparent difference is a vertical shift by a constant over all of $x \in [0, 10]$. A data set of 200 points, plotted in green in Figure 4.7, is generated following a true mean function of $m(x) = -0.125(x-5)^3 + x$ with a true variance function of $v(x) = \exp(-(x-5)^2/5)$. A second set of data (plotted in purple in Figure 4.7) is simulated from the mean function but also shifted by a constant $S = 10$.

We consider the following semiparametric model to describe this data:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 S_i + \sum_{j=1}^{K} b_j (x_i - \kappa_j)_+^2 + \epsilon_i \\ \log(\sigma_i^2) &= \delta_0 + \delta_1 x_i + \delta_2 x_i^2 + \sum_{k=1}^{K_v} c_k (x_i - \kappa_{v_k})_+^2 . \end{aligned} \tag{87}$$

Here S_i is parametric covariate that takes value 10 if x_i comes from the shifted data and 0 otherwise. Functionally, this model is equivalent to

$$\begin{aligned} y_i &= \beta_3 S_i + f(x_i) + \epsilon_i \\ \log(\sigma_i^2) &= g(x_i). \end{aligned} \tag{88}$$

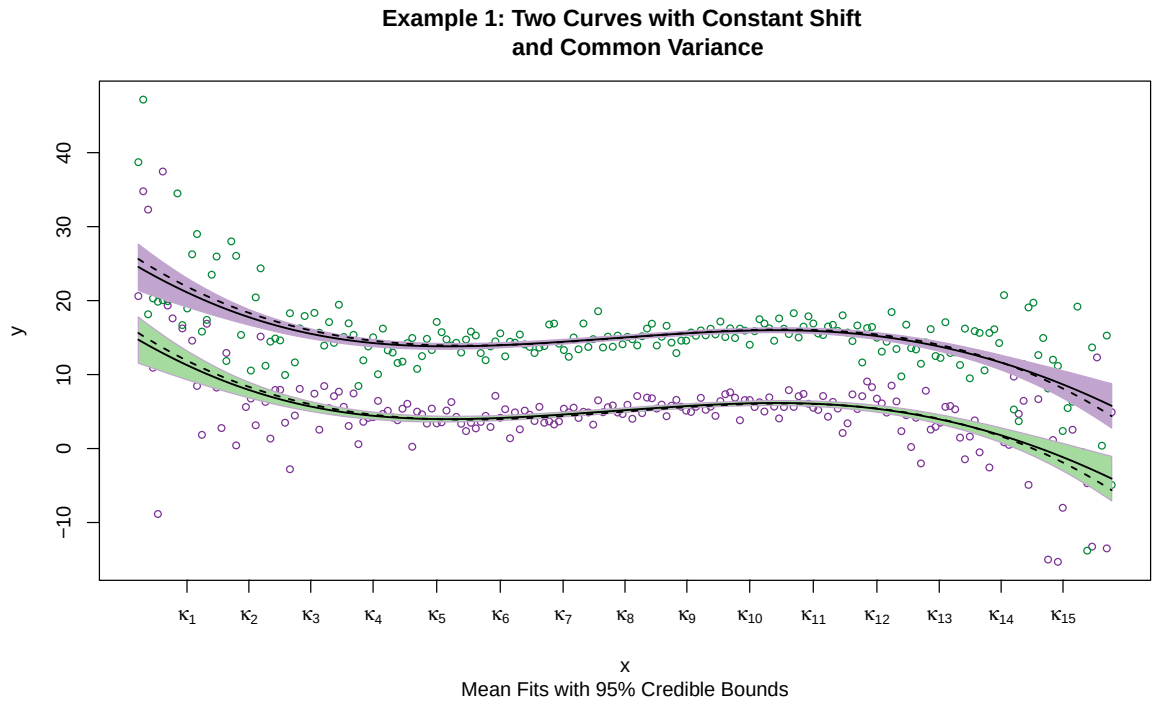
Since this model only has a single nonparametric component for each level ($f(x)$ and $g(x)$), $L = M = 1$. The only difference between this model and the single curve cases considered earlier is the inclusion of the shift variable S_i, which adds an additional column to the fixed effect matrix $\mathbf{X}$. While there is noticeable heteroskedasticity in the data, there does not appear to be a difference in variance structures of the data sets so the assumption of a single log variance function $g(x)$ is appropriate.

4.3.9 Semiparametric Interaction Model with Common Variance Structure

For this example the first data set (purple in Figure 4.8) is generated following a true mean function of $m_1(x) = \exp(-x^2/12)$. The second data set (green in Figure 4.8) is generated from a related mean curve of the form $m_2(x) = \exp(-x^2/(12 + S))$ where the shift constant $S = 8$. Both data sets are generated according to a true variance function $v(x) = (x/100)^2$.

Data of this type can typically be modeled by including a parametric-by-nonparametric interaction term in the mean level if one has knowledge of some parametric covariate that is related to the change between curves. Here, we let S_i be an observed covariate that takes the value 0 if y_i is generated from $m_1(x)$ and 8 if the response is generated from $m_2(x)$. Functionally, the semiparametric interaction model would be

$$\begin{aligned} y_i &= f_1(x_i) + S_i * f_2(x_i) + \epsilon_i \\ \log(\sigma_i^2) &= g(x_i). \end{aligned} \tag{89}$$



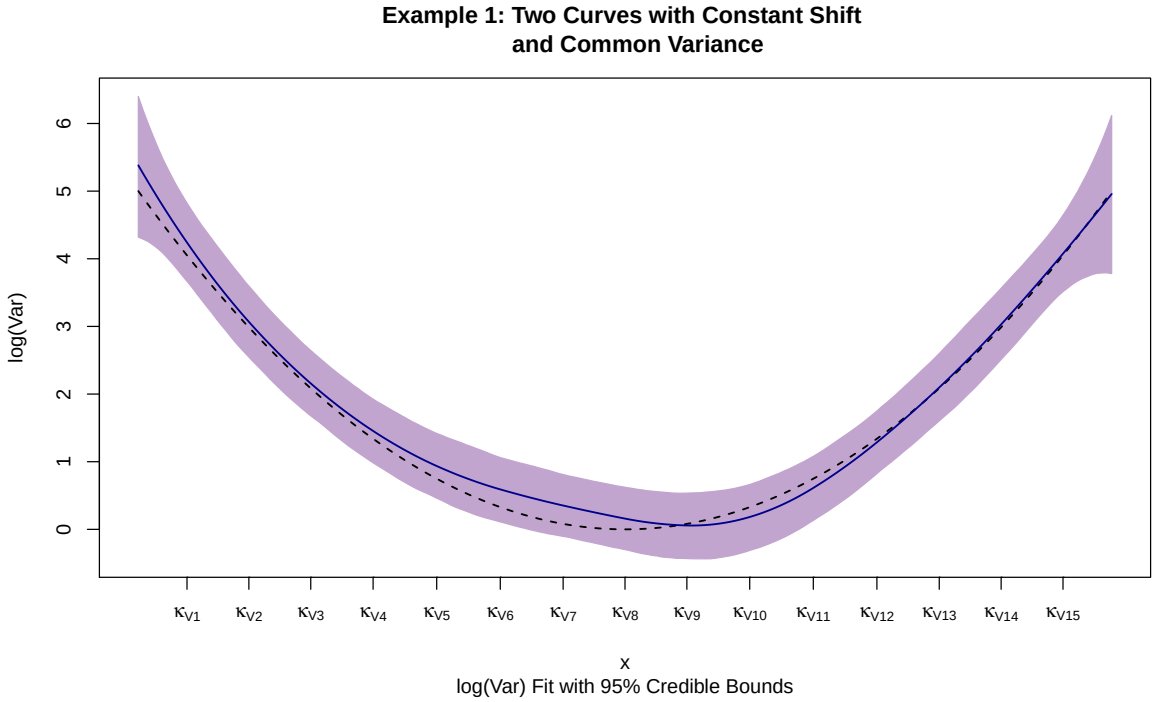


Figure 4.7: Variational approximation of parametric shift model with common variance. Green points represent data simulated from a true mean function $m_1(x) = -.125(x-5)^3 + x$ and purple points represent data simulated by a shifted true mean function $m_2(x) + 10$. The true variance function for both curves is $v(x) = \exp\left(-(x-5)^2/5\right)$.

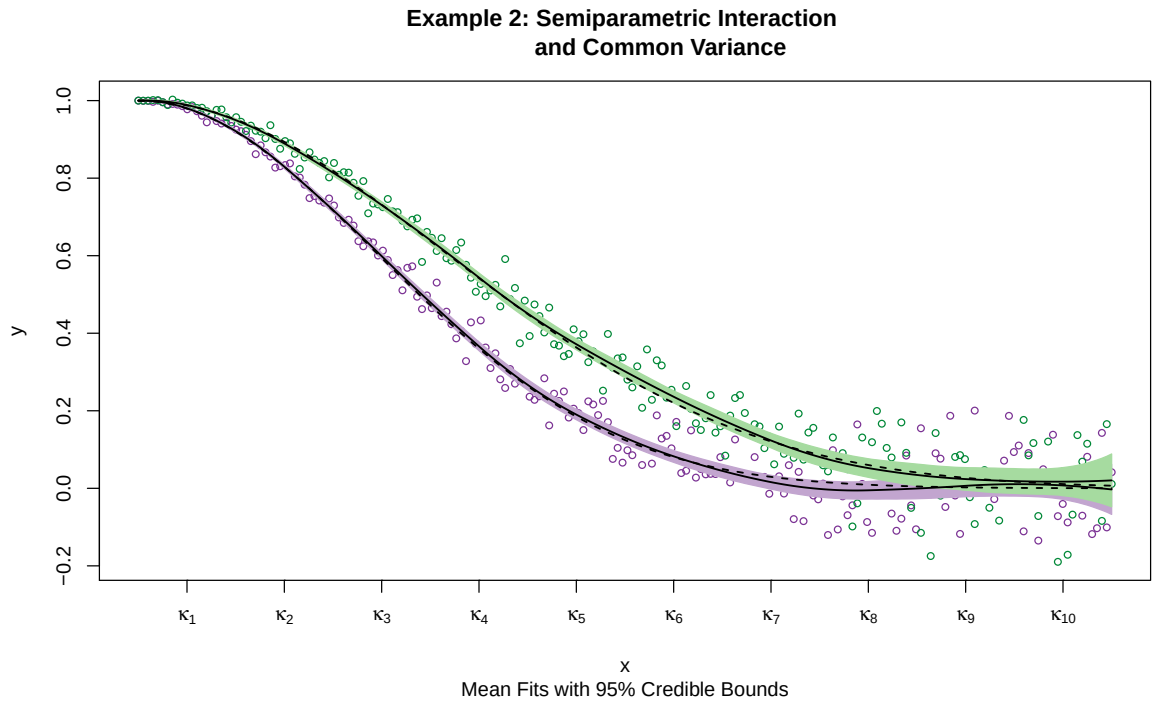
Using our truncated quadratic basis, this model has the form

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 S_i + \beta_4 x_i S_i + \beta_5 x_i^2 S_i \\ &\quad + \sum_{j=1}^{K} b_{1j}(x_i - \kappa_j)_+^2 + \sum_{j=1}^{K} S_i b_{2j}(x_i - \kappa_j)_+^2 + \epsilon_i \\ \log(\sigma_i^2) &= \delta_0 + \delta_1 x_i + \delta_2 x_i^2 + \sum_{k=1}^{K_V} c_k(x_i - \kappa_k)_+^2. \end{aligned} \tag{90}$$

Under this formulation, it is apparent that the example described in Section 4.3.8 uses a submodel of this semiparametric interaction model. The model in (90) contains six fixed effect terms and two sets of 15 random effect terms. The random effects $b_{11}, \ldots, b_{1,15}$ are associated with the function $f_1(x)$ while the random effects $b_{2,1}, \ldots, b_{2,15}$ are associated with the interaction function $f_2(x)$. The introduction of an interaction term requires two separate random effect variance terms for the model, $\sigma_{b_1}^2$ and $\sigma_{b_2}^2$ ($L = 2$). As with the example in Section 4.3.8, there is no evidence of differing variance structures between the two data sets so assuming a common log-variance function $g(x)$ is founded.

4.3.10 Semiparametric Interaction Model with Dampening Variance

This example is an extension of the one described in Section 4.3.9. Figure 4.9 shows the simulated data and the resulting variational approximation. Two sets of data are generated from mean functions $m_1(x)$ and $m_2(x)$ with shift parameter $S = 8$. However, the data sets generated in this example have different variance functions. The variance function associated with $m_1(x)$ (purple in Figure 4.9) is $v_1(x) = (x/100)^2$, the same as Section 4.3.9. The data simulated from $m_2(x)$ follows a damped version of this variance function defined as $v_2(x) = 0.05(x/100)^2$.



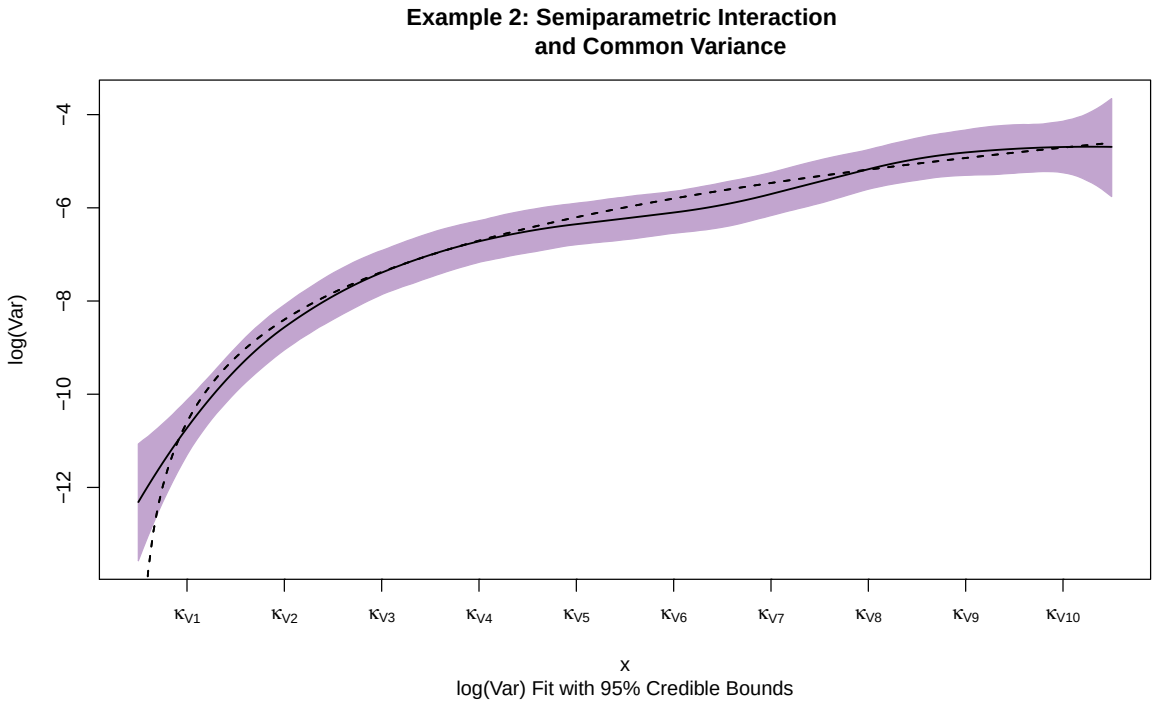


Figure 4.8: Variational approximation of semiparametric interaction model with common variance. Purple points represent data simulated from a true mean function $m_1(x) = \exp(-x^2/12)$ and green points represent data simulated from $m_2(x) = \exp(-x^2/(12+S))$ with shift parameter $S = 8$. The true variance function for both curves is $v(x) = (x/100)^2$.

One possible model for this data is to extend (90) to include a semiparametric interaction at the variance level as well. Functionally, this takes the form

$$\begin{aligned} y_i &= f_1(x_i) + S_i * f_2(x_i) + \epsilon_i \\ \log(\sigma_i^2) &= g_1(x_i) + D_i * g_2(x_i). \end{aligned} \tag{91}$$

The shift covariate S_i is defined as before and D_i is an indicator variable such that

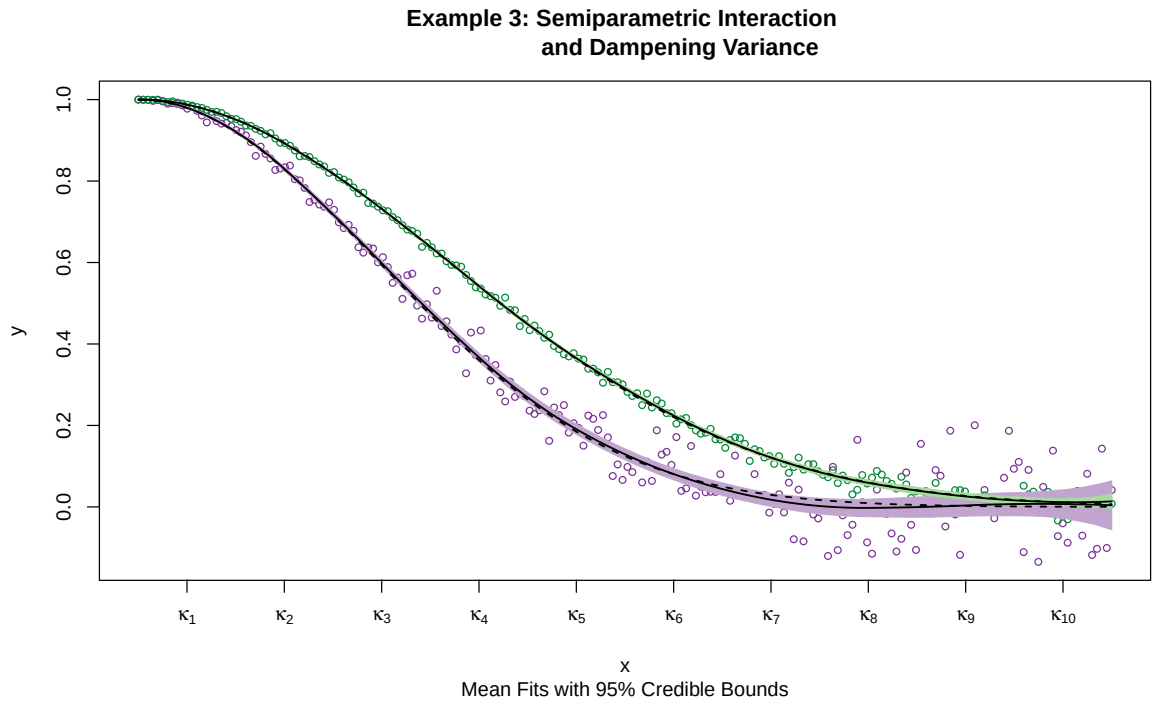
$$D_i = \begin{cases} 0 & \text{if generated from } m_1(x_i) \\ 1 & \text{if generated from } m_2(x_i). \end{cases} \tag{92}$$

Using our truncated quadratic basis, (91) is expressed as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 S_i + \beta_4 x_i S_i + \beta_5 x_i^2 S_i \\ &\quad + \sum_{j=1}^{K} b_{1j}(x_i - \kappa_j)_+^2 + \sum_{j=1}^{K} S_i b_{2j}(x_i - \kappa_j)_+^2 + \epsilon_i \\ \log(\sigma_i^2) &= \delta_0 + \delta_1 x_i + \delta_2 x_i^2 + \delta_3 D_i + \delta_4 x_i D_i + \delta_5 x_i^2 D_i \\ &\quad + \sum_{k=1}^{K_V} c_{1k}(x_i - \kappa_{V_k})_+^2 + \sum_{k=1}^{K_V} D_i c_{2k}(x_i - \kappa_{V_k})_+^2. \end{aligned} \tag{93}$$

4.3.11 Computational Performance of Multiple Curve Models

In Table 4.2, we examine the run times of the three multiple curve examples presented above. For comparison, run times from the hybrid Gibbs-DRAM MCMC method (Section 3.2) are included in Table 4.2 as well. Increases in model complexity, particularly the inclusion of interaction terms at both the mean and variance levels affect the computation time for both variational approximation and the MCMC method, as one would expect. In all cases, the variational approximation still considerably outpaces the MCMC method, with



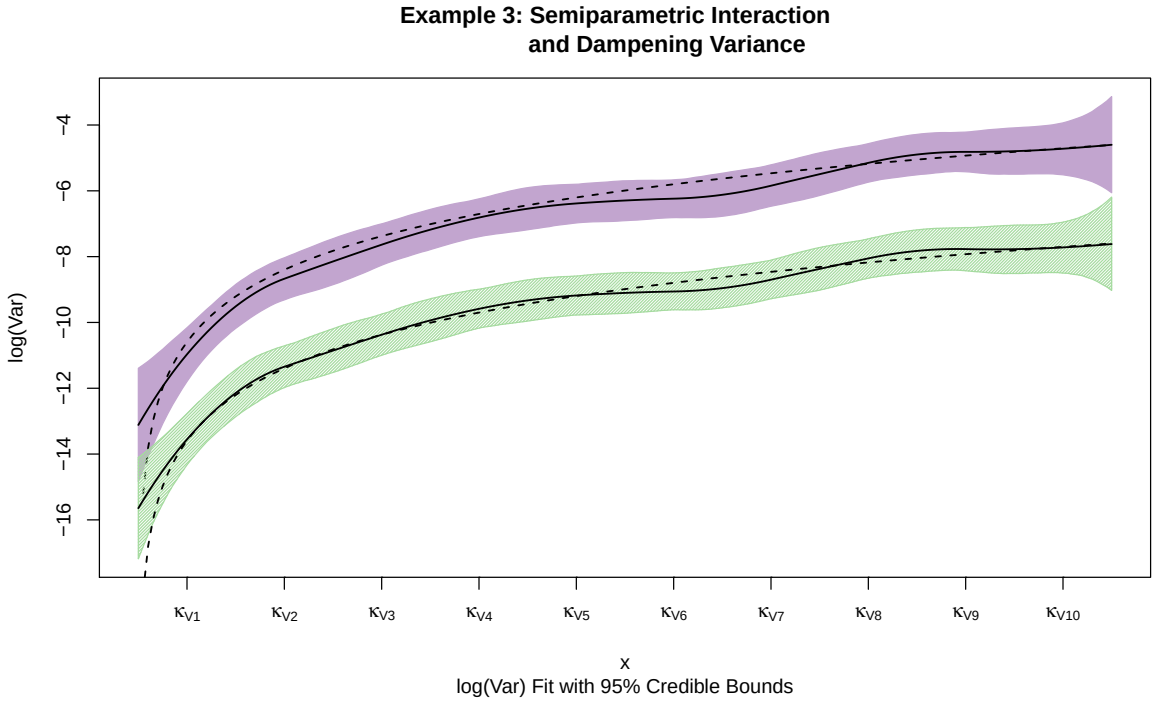


Figure 4.9: Variational approximation of semiparametric interaction model with damped variance. Purple points represent data simulated from a true mean function $m_1(x) = \exp(-x^2/12)$ and green points represent data simulated from $m_2(x) = \exp(-x^2/(12+S))$ with shift parameter $S = 8$. The true variance functions associated with $m_1(x)$ and $m_2(x)$ are $v_1(x) = (x/100)^2$ and $v_2(x) = 0.05(x/100)^2$ respectively.

improvements ranging from 3 to 10 times faster computation depending on sample size and model complexity. It should be remembered that the number of total parameters in even relatively simple models grow rapidly as interaction terms are added. For reference, the $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_V$ parameters in the semiparametric interaction model with dampening variance are both 33-dimensional parameter vectors. Additional model complexity would highlight the usefulness of these variational approximations, particularly if the MCMC procedure required more than $10,000$ iterations to reach an appropriate sampling of the posterior parameter distribution.

Table 4.2: Run times (seconds) for the variational approximation described in Algorithm 3 for the constant shift model (Model 1), semiparametric interaction model with common variance (Model 2), and the semiparametric interaction model with dampening variance (Model 3) examples. For comparison, the run-times for a hybrid Gibbs-DRAM MCMC method are included akin to the results found in Table 4.1. The MCMC methods were ran for $10,000$ iterations with a burn-in of 1000. Sample sizes correspond to total number of simulated values split equally across curves.

Method	Sample Size (N)			
	200	400	800	1600
Variational approximation for Model 1	18.64	37.11	63.50	161.27
Variational approximation for Model 2	15.76	29.37	60.34	136.99
Variational approximation for Model 3	29.17	55.96	118.47	305.10
MCMC for Model 1	64.19	140.04	404.83	1641.53
MCMC for Model 2	63.32	157.54	550.19	1926.43
MCMC for Model 3	84.47	183.69	581.16	2186.63

4.4 Laplace Variational Approximation for Joint Mean-Variance Model

Following the derivation of variational Bayes methods for the first two models presented in Chapter 2, we now present preliminary work on implementing a variational approximation for the joint mean-variance (JMV) model motivated by the radial reduction process that occurs in SAXS experiments. For illustrative convenience, we limit our focus to the simple

radial variance fixed effect model presented in (40). In this section we detail our first efforts on this problem, particularly focusing on the difficulties arising from the relationship between the radial and model variances.

Recall the posterior parameter conditionals for the radial fixed effects JMV model, rewritten in (94) for convenience.

$$\begin{aligned}
\boldsymbol{\theta} \mid \cdot &\sim \mathcal{N}\left(\mathbf{M}\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}, \mathbf{M}\right) \text{ where } \mathbf{M} = \left({\boldsymbol{\Sigma}_\theta}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}\right)^{-1} \\
\boldsymbol{\eta} \mid \cdot &\sim \mathcal{N}\left(\frac{1}{\sigma_u^2}\mathbf{M}_W\mathbf{X}_W^T\mathbf{w}, \mathbf{M}_W\right) \text{ where } \mathbf{M}_W = \left({\boldsymbol{\Sigma}_\eta}^{-1} + \frac{1}{\sigma_u^2}\mathbf{X}_W^T\mathbf{X}_W\right)^{-1} \\
\sigma_b^2 \mid \cdot &\sim \mathcal{IG}\left(A_b + \frac{K}{2}, B_b + \frac{\|\mathbf{b}\|^2}{2}\right) \\
\sigma_c^2 \mid \cdot &\sim \mathcal{IG}\left(A_c + \frac{K_V}{2}, B_c + \frac{\|\mathbf{c}\|^2}{2}\right) \\
p(\boldsymbol{\theta}_V \mid \cdot) &\propto \exp\Big[-\frac{1}{2}\Big\{\sum_{i=1}^{N}{\mathbf{C}_{\mathbf{V_i}}}^T\theta_{\mathbf{V}} + \sum_{i=1}^{N}\left(Y_i - {\mathbf{C}_{\mathbf{i}}}^T\theta\right)^2\exp(-{\mathbf{C}_{\mathbf{V_i}}}^T\theta_{\mathbf{V}}) \\
&\qquad + {\theta_{\mathbf{V}}}^T{\boldsymbol{\Sigma}_{\theta_{\mathbf{V}}}}^{-1}\theta_{\mathbf{V}} + \frac{1}{\sigma^2}\|\mathbf{w} - \mathbf{X}_W\boldsymbol{\eta}\|^2\Big\}\Big] \\
\sigma_u^2 \mid \cdot &\sim \mathcal{IG}\left(A_u s + \frac{N}{2}, B_u + \frac{\|\mathbf{w} - \mathbf{X}_W\boldsymbol{\eta}\|^2}{2}\right).
\end{aligned} \tag{94}$$

A running theme throughout this dissertation is the similarities between the JMV model and the heteroskedastic semiparametric regression model. As such, our first attempt at a variational approximation is based on an extension of the method detailed in Section 4.2. Let $\chi = (\boldsymbol{\theta}, \boldsymbol{\theta}_V, \boldsymbol{\eta}, \sigma_b^2, \sigma_c^2, \sigma_u^2)$. Assume the variational density $q(\chi)$ has the product density form

$$q(\chi) = q_1(\theta)q_2(\boldsymbol{\theta}_V)q_3(\boldsymbol{\eta})q_4(\sigma_b^2)q_4(\sigma_c^2)q_5(\sigma_u^2). \tag{95}$$

From this assumption, one can use (57) to determine the optimal variational density of a particular parameter based on the variational expectation of the log posterior parameter

conditional. As before, the numbered notation on the variational densities will be dropped for convenience (e.g. $q_1(\theta) = q(\theta)$).

The derivations of the optimal variational densities for θ, σ_b^2, and σ_c^2 are simple and follow from the same methods used in Section 4.3. These densities are

$$\begin{aligned}
\boldsymbol{\theta} &\overset{q^*}{\sim} \mathcal{N}\left(\mu_{q(\boldsymbol{\theta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\right) \\
\sigma_b^2 &\overset{q^*}{\sim} \mathcal{IG}\left(A_b + \frac{K}{2}, B_{q(\sigma_b^2)}\right) \\
\sigma_c^2 &\overset{q^*}{\sim} \mathcal{IG}\left(A_c + \frac{K_V}{2}, B_{q(\sigma_c^2)}\right),
\end{aligned} \tag{96}$$

with variational parameters

$$\begin{aligned}
\boldsymbol{\Sigma}_{q(\theta)} &= \left(\text{blockdiag}(\frac{1}{\sigma_\beta^2}\mathcal{I}_p, \mu_{q(1/\sigma_{b_1}^2)}\mathcal{I}_{K_1}, \ldots, \mu_{q(1/\sigma_{b_L}^2)}\mathcal{I}_{K_L}) + \mathbf{C^T \boldsymbol{\Gamma} C}\right)^{-1} \\
\boldsymbol{\mu}_{q(\theta)} &= \boldsymbol{\Sigma}_{\mathbf{q}(\boldsymbol{\theta})}\mathbf{C^T \boldsymbol{\Gamma} y} \\
B_{q(\sigma_b^2)} &= B_b + \frac{1}{2}\left(\mu_{q(\mathbf{b})}^2 + \text{trace}(\boldsymbol{\Sigma}_{q(\mathbf{b})})\right) \\
B_{q(\sigma_c^2)} &= B_c + \frac{1}{2}\left(\mu_{q(\mathbf{c})}^2 + \text{trace}(\boldsymbol{\Sigma}_{q(\mathbf{c})})\right)
\end{aligned} \tag{97}$$

where

$$\boldsymbol{\Gamma} = \text{diag}\Big(E_{-\boldsymbol{\theta}}\left[\exp\left\{-\mathbf{C}_{V_1}^T\boldsymbol{\theta}_V\right\}\right], \ldots, E_{-\boldsymbol{\theta}}\left[\exp\left\{-\mathbf{C}_{V_N}^T\boldsymbol{\theta}_V\right\}\right]\Big). \tag{98}$$

The elements of $\boldsymbol{\Gamma}$ correspond to the moment generating function associated with $q(\boldsymbol{\theta}_V)$ evaluated at the vector $-\mathbf{C}_{V_i}^T$ for all $i = 1, \ldots, N$.

The derivation of the optimal variational density for the regression coefficients of the radial variance model, $\boldsymbol{\eta}$, is more complex than that of $\boldsymbol{\theta}$, despite having a similar posterior conditional structure. This is because of the use of the log model variance as a covariate in

the fixed effect design matrix X_W. Consider

$$E_{-\boldsymbol{\eta}}\left[\log p(\boldsymbol{\eta} \mid \cdot)\right] = E_{-\boldsymbol{\eta}}\Big[-\frac{N}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{M}_W|) \\ - \frac{1}{2}\left(\boldsymbol{\eta} - \frac{1}{\sigma_u^2}\mathbf{M}_W\mathbf{X}_W^T\mathbf{w}\right)^T \mathbf{M}_W^{-1}\left(\boldsymbol{\eta} - \frac{1}{\sigma_u^2}\mathbf{M}_W\mathbf{X}_W^T\mathbf{w}\right)\Big] \tag{99}$$

The first two terms inside the expectation can be dropped for now since we are ultimately interested in the proportional form of $q(\boldsymbol{\eta})$. Since the log model variance, which is directly a function of $\boldsymbol{\theta}_V$, is a covariate in $\mathbf{X}_W$, explicit knowledge of the model structure is a prerequisite for any variational approximation implementation. That is, the form of $\mathbf{X}_W$, particularly how v is included, is required for deriving the forms of the variational parameters of $q(\boldsymbol{\eta})$. This is the main reason why we limit our initial investigation to the fixed effect radial variance model.

Since the third term in (99) is quadratic with respect to $\boldsymbol{\eta}$ regardless of choice of $\mathbf{X}_W$, the optimal variational density for $\boldsymbol{\eta}$ will be Gaussian with variational mean $\mu_{q(\eta)}$ and covariance matrix $\boldsymbol{\Sigma}_{q(\eta)}$. Under the fixed effect radial variance model, these variational parameters are

$$\begin{aligned} \boldsymbol{\Sigma}_{q(\boldsymbol{\eta})} &= \left(\mu_{q(1/\sigma_u^2)}\mathbf{Q} + \frac{1}{\sigma_\eta^2}\mathcal{I}_3\right)^{-1} \\ \mu_{q(\boldsymbol{\eta})} &= \mu_{q(1/\sigma_u^2)}\boldsymbol{\Sigma}_{q(\boldsymbol{\eta})}^T\boldsymbol{\Psi}^T\mathbf{w}, \end{aligned} \tag{100}$$

where the matrix $\mathbf{Q}$ is

$$\mathbf{Q} = \left(\mu_{q(\boldsymbol{\theta}_V)}^T\mathbf{X}_W^T\mathbf{X}_W\mu_{q(\boldsymbol{\theta}_V)} + \sum_{i=1}^{N}\mathbf{C}_{V_i}^T\boldsymbol{\Sigma}_{q(\boldsymbol{\theta}_V)}\mathbf{C}_{V_i}\right) \tag{101}$$

and $\boldsymbol{\Psi}$ is the $N \times 3$ matrix $\left[1, \mathbf{x}, \mathbf{C}_V\mu_{q(\boldsymbol{\theta}_V)}\right]$.

The derivation of $q(\sigma_u^2)$ also depends on the radial variance model structure. Using the posterior conditional from (94), the optimal variational density for σ_u^2 corresponds to an

$\mathcal{IG}\left(A_u + N/2, B_{q(\sigma_u^2)}\right)$. The variational parameter $B_{q(\sigma_u^2)}$ is

$$B_{q(\sigma_u^2)} = B_u + \frac{1}{2} E_{-\sigma_u^2}\left[\|\mathbf{w} - \mathbf{X}_W \boldsymbol{\eta}\|^2\right]. \tag{102}$$

The expectation term here can be written in terms of the variational parameters for $q(\boldsymbol{\theta}_V)$ and $q(\boldsymbol{\eta})$.

The nonconjugacy of the posterior conditional distribution $\boldsymbol{\theta}_V \mid \cdot$ yields the variational density

$$\begin{aligned} q^*(\boldsymbol{\theta}_V) \propto \exp\Bigg\{ -\frac{1}{2}\Bigg(& \sum_{i=1}^{N} \left[(y_i - \mathbf{C}_i^T \mu_{q(\boldsymbol{\theta})})^2 + \mathbf{C}_i^T \boldsymbol{\Sigma}_{\mathbf{q}(\boldsymbol{\theta})}{}^T \mathbf{C}_i\right] \exp\{-\mathbf{C}_{V_i}^T \boldsymbol{\theta}_V\} \\ & + \sum_{i=1}^{N} \mathbf{C}_{V_i}^T \boldsymbol{\theta}_V + \boldsymbol{\theta}_V{}^T \boldsymbol{\Omega} \boldsymbol{\theta}_V + \mu_{q(1/\sigma_u^2)} E_{-\boldsymbol{\theta}_V}\left[\|\mathbf{w} - \mathbf{X}_W \eta\|^2\right] \Bigg) \Bigg\}, \end{aligned} \tag{103}$$

where $\boldsymbol{\Omega} = \text{blockdiag}\left(\sigma_\delta^{-2}\mathcal{I}_r, \sigma_c^{-2}\mathcal{I}_{K_V}\right)$. Since the last term is the expectation with respect to all parameters save $\boldsymbol{\theta}_V$, this can be written as

$$E_{-\boldsymbol{\theta}_V}\left[\|\mathbf{w} - \mathbf{X}_W \eta\|^2\right] = \left\|\mathbf{w} - \mathbf{X}_W \mu_{q(\eta)}\right\|^2 + \sum_{i=1}^{N} \mathbf{X}_{W_i}^T \boldsymbol{\Sigma}_{q(\eta)} \mathbf{X}_{W_i}. \tag{104}$$

Following from Section 4.3.1, we can use a Laplace approximation approach to approximate this variational density with a Gaussian distribution denoted $\tilde{q}(\boldsymbol{\theta}_V)$. Let $h(\boldsymbol{\theta}_V \mid \cdot)$ be the function such that $q^*(\boldsymbol{\theta}_V) \propto \exp\{-h(\boldsymbol{\theta}_V \mid \cdot)\}$ and $\alpha = \text{argmin}\, h(\boldsymbol{\theta}_V \mid \cdot)$. Recall that under the Laplace approximation, the variational parameters for $q(\boldsymbol{\theta}_V)$ are defined as

$$\begin{aligned} \mu_{\tilde{q}(\boldsymbol{\theta}_V)} &= \alpha \\ \boldsymbol{\Sigma}_{\tilde{q}(\boldsymbol{\theta}_V)} &= \left(\frac{\mathbf{H}_{\boldsymbol{\theta}_V}(\alpha)}{2}\right)^{-1}, \end{aligned} \tag{105}$$

where $\mathbf{H}_{\boldsymbol{\theta}_V}(\alpha)$ is the Hessian matrix of $h(\boldsymbol{\theta}_V \mid \cdot)$ evaluated at α.

Algorithm 4 details an iterative approach to determining the appropriate variational parameters for this approximation. As with previous methods, convergence is monitored through successive values of $\log \underline{p}(\mathbf{y}; q)$, as defined in (55). Once converged, the final parameter estimates are taken as the mean of their corresponding variational densities.

Algorithm 4 Iterative method for determining the optimal variational distributions for the parameters of the joint radial mean-variance model with fixed effect radial variance structure.

Initialize: $\mu_{q(\boldsymbol{\theta})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$, $\mu_{q(\boldsymbol{\theta}_V)}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta}_V)}$, $B_{q(\sigma_b^2)}$,$B_{q(\sigma_c^2)}$, $B_{q(\sigma_u^2)}$

repeat

$\mathbf{Q} \leftarrow \left(\mu_{q(\boldsymbol{\theta}_V)}^T \mathbf{X}_W^T \mathbf{X}_W \mu_{q(\boldsymbol{\theta}_V)} + \sum_{i=1}^N \mathbf{C}_{V_i}^T \boldsymbol{\Sigma}_{q(\boldsymbol{\theta}_V)} \mathbf{C}_{V_i}\right)$

$\boldsymbol{\Sigma}_{q(\boldsymbol{\eta})} \leftarrow \left(\mu_{q(1/\sigma_u^2)} \mathbf{Q} + \frac{1}{\sigma_\eta^2} \mathcal{I}_3\right)^{-1}$

$\boldsymbol{\Psi} \leftarrow \left[1, \mathbf{x}, \mathbf{C}_V \mu_{q(\boldsymbol{\theta}_V)}\right]$

$\mu_{q(\boldsymbol{\eta})} \leftarrow \mu_{q(1/\sigma_u^2)} \boldsymbol{\Sigma}_{q(\boldsymbol{\eta})}^T \boldsymbol{\Psi}^T \mathbf{w}$

$B_{q(\sigma_u^2)} \leftarrow B_u + \frac{1}{2} E_{-\sigma_u^2}\left[\|\mathbf{w} - \mathbf{X}_W \boldsymbol{\eta}\|^2\right]$

Compute $\alpha = \text{argmin } h(\boldsymbol{\theta}_V \mid \cdot)$, $\mathbf{H}_{\boldsymbol{\theta}_V}(\alpha)$

$\mu_{\tilde{q}(\boldsymbol{\theta}_V)} \leftarrow \alpha$

$\boldsymbol{\Sigma}_{\tilde{q}(\boldsymbol{\theta}_V)} \leftarrow \left(\frac{\mathbf{H}_{\boldsymbol{\theta}_V}(\alpha)}{2}\right)^{-1}$

$B_{q(\sigma_c^2)} \leftarrow B_c + \frac{1}{2}\|\mu_{q(\mathbf{c})}\|^2 + \frac{1}{2}\text{trace}(\boldsymbol{\Sigma}_{q(\mathbf{c})})$

$\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \leftarrow \left(\text{blockdiag}(\frac{1}{\sigma_\beta^2}\mathcal{I}_p, \mu_{q(1/\sigma_b^2)}\mathcal{I}_K) + \mathbf{C}^\mathbf{T}\boldsymbol{\Gamma}\mathbf{C}\right)^{-1}$

$\mu_{q(\boldsymbol{\theta})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \mathbf{C}^T \boldsymbol{\Gamma} \mathbf{y}$

$B_{q(\sigma_b^2)} \leftarrow B_b + \frac{1}{2}\left(\mu_{q(\mathbf{b})}^2 + \text{trace}(\boldsymbol{\Sigma}_{q(\mathbf{b})})\right)$

until Convergence is Reached

Construct parameter estimates using mean of variational distributions

4.4.1 Simulation Example

Consider the simulated radial mean and variance data from Section 3.3.2. Algorithm **??** is used to fit the data, displayed in Figure 4.10. The top panel shows that the approximate radial mean fit appropriately describes the underlying true mean as well as reflecting the

heteroskedasticity of the simulated data. The bottom panel shows a very good fit of the simulated radial variance data generated under the fixed effect model $w = 10 - 4\log(x) + v$.

Figure 4.11 depicts the approximate estimate of the log model variance function. While the uncertainty bounds cover the true log model variance, the estimated fit appears to have difficulty matching the concavity of the true function. This may partially be explained as an artifact of the embedded Gaussian approximation used to estimate the variational density $q^*(\boldsymbol{\theta}_V)$. However further investigation indicates that the "naive" product density assumption made in (95) may lead to inappropriate model variance estimates. This problem hinges on the relationship between $\boldsymbol{\theta}_V$ and $\boldsymbol{\eta}$. Under the current product density for $q(\chi)$, the assumption that $q(\boldsymbol{\theta}_V, \boldsymbol{\eta}) = q(\boldsymbol{\theta}_V)q(\boldsymbol{\eta})$ may in fact be too strong. This could lead to false inference in cases where the relative importance of the log model variance to the radial variance model is particularly high.

To explore this effect, the simulated data above is modified to have two different generating radial variance functions. The modified data, denoted w_2 and w_3 are

$$\begin{aligned} w_{2_i} &= 10 - 4\log(x_i) + 10v_i + u_i \\ w_{3_i} &= 10 - 10\log(x_i) + v_i \\ u_i &\sim \mathcal{N}\left(0, \sigma_u^2\right), \end{aligned} \tag{106}$$

with $\sigma_u^2 = 0.25$ as before. Figures 4.12 and 4.13 show the variational approximation for the w_2 case. This is a case where the influence of v over the log radial variance has been increased relative to the influence of $\log(x)$. Most surprisingly, Figure 4.13 indicates that this approximation completely misses the mark for describing the log model variance function for small values of x. This overestimation leads to incorrect uncertainty bounds at the radial mean level.

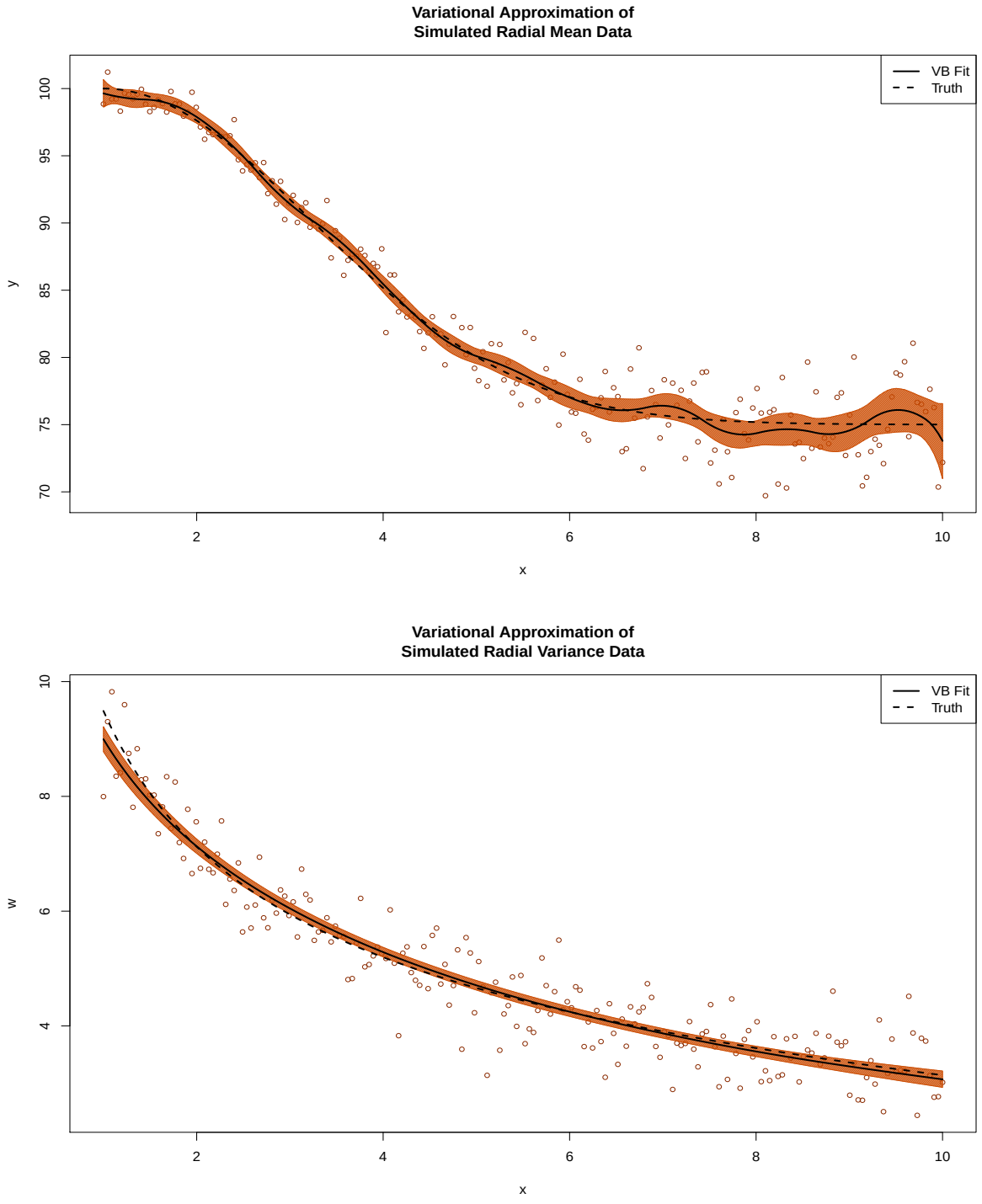


Figure 4.10: Approximate fits for the simulated radial mean and variance data using the variational method described in Algorithm 4. The shaded regions repesent 95% credible regions using the converged variational parameters for $q(\theta)$ and $q(\eta)$.

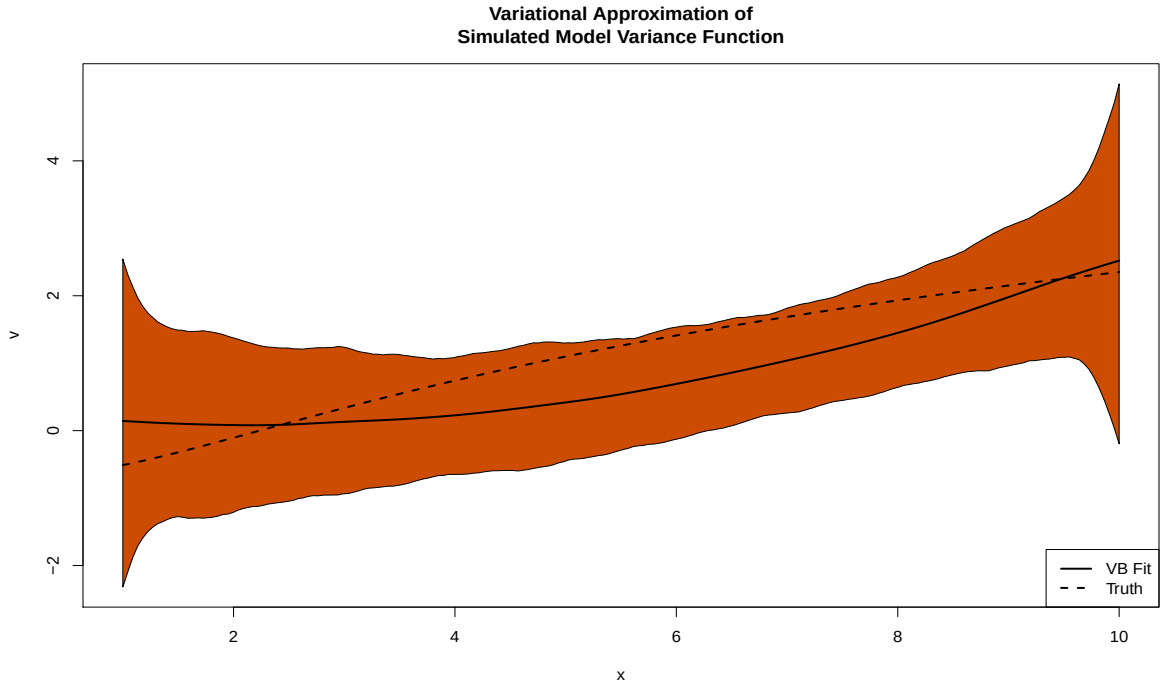


Figure 4.11: Approximate fit of the log model variance function for the simulated radial mean and variance data using Algorithm 4. The shaded region represents a 95% credible region using the converged variational parameters of $\tilde{q}(\boldsymbol{\theta}_V)$

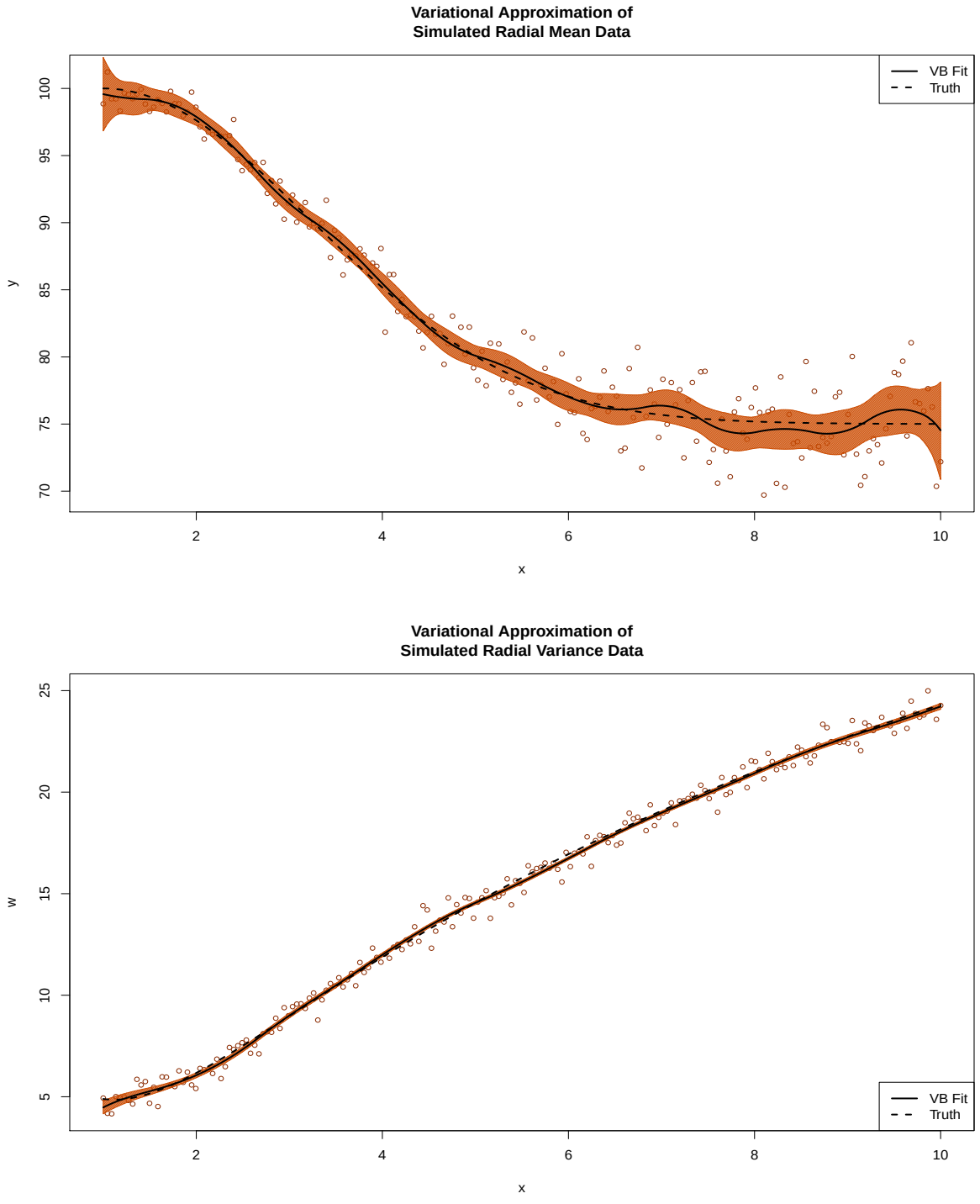


Figure 4.12: Approximate fits for the simulated radial mean and variance data when the radial variance function is $w_2(x, v)$. The shaded regions repesent 95% credible regions using the converged variational parameters for $q(\theta)$ and $q(\eta)$.

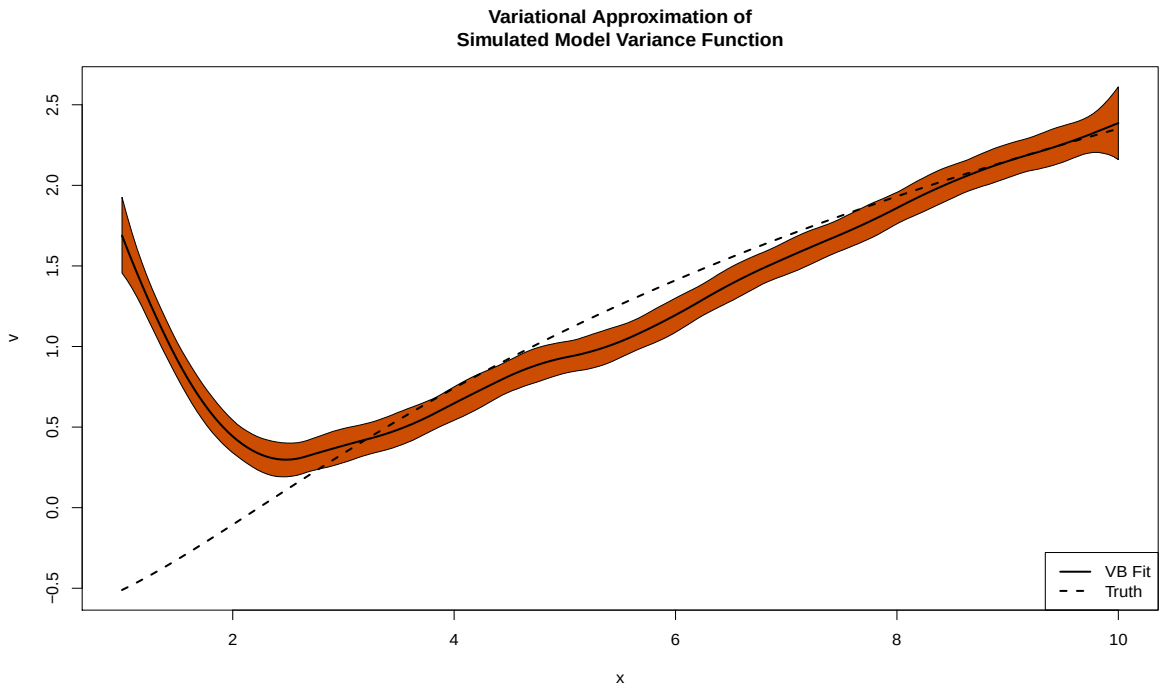


Figure 4.13: Approximate fit of the log model variance function when the radial variance function is $w_2(x, v)$. The shaded region represents a 95% credible region using the converged variational parameters of $\tilde{q}(\boldsymbol{\theta}_V)$

Figures 4.14 and 4.15 show the variational approximation for the case where the relative impact of the log model variance function is decreased. Under this scenario, we observe behavior much like the original simulation example. The radial mean data is well fit and comparable to the original simulation example. The radial variance data is fit as before, but with significantly tighter uncertainty bounds. Figure 4.15 is virtually indistinguishable from the estimate in the original simulation example.

The anecdotal evidence presented with these three simulation examples suggests that simply extending the variational approximation of the heteroskedastic semiparametric regression model, detailed in Algorithm 3, may not be sufficient. A potential alternative would be to assume the variational density $q(\chi)$ expands to

$$q(\chi) = q(\boldsymbol{\theta})q(\sigma_b^2)q(\sigma_c^2)q(\sigma_u^2)q(\boldsymbol{\eta}, \boldsymbol{\theta}_V). \tag{107}$$

Here the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\theta}_V$ would be approximated jointly. Most likely this would result in a nonconjugate form similar to that of $q(\boldsymbol{\theta}_V)$ in (103). Future work will focus on the implementation of a variational approximation using this product density structure, particularly on the appropriateness of imposing a Gaussian form on the noncojugate variational density by using the Laplace approximation method described throughout this work.

4.4.2 Extensions to Semiparametric Radial Variance Models

The variational approximation presented in Algorithm 4 is specifically limited to models where the radial variance data is a three-term fixed effect model. This is primarily done to ease derivation of the optimal q densities. If one were to consider models where the log model variance, v, is treated nonparametrically, the form of the basis functions must be considered. For example, if a truncated polynomial basis of degree p is used, the variational expectations

$$E\left[(v_i - \kappa_k^V)_+^j\right] \quad \forall j = 1, \ldots, 2p \tag{108}$$

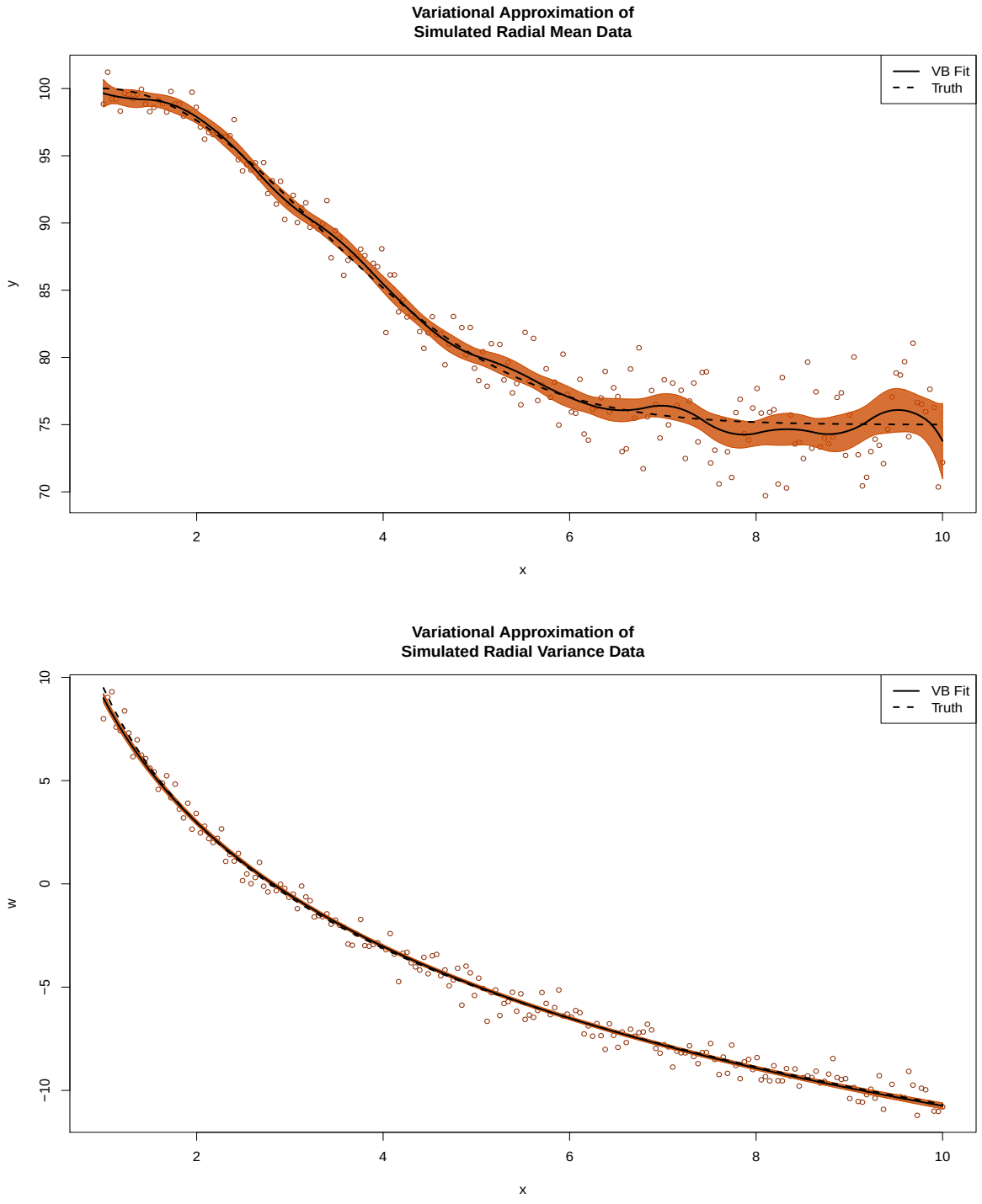


Figure 4.14: Approximate fits for the simulated radial mean and variance data when the radial variance function is $w_2(x, v)$. The shaded regions repesent 95% credible regions using the converged variational parameters for $q(\theta)$ and $q(\eta)$.

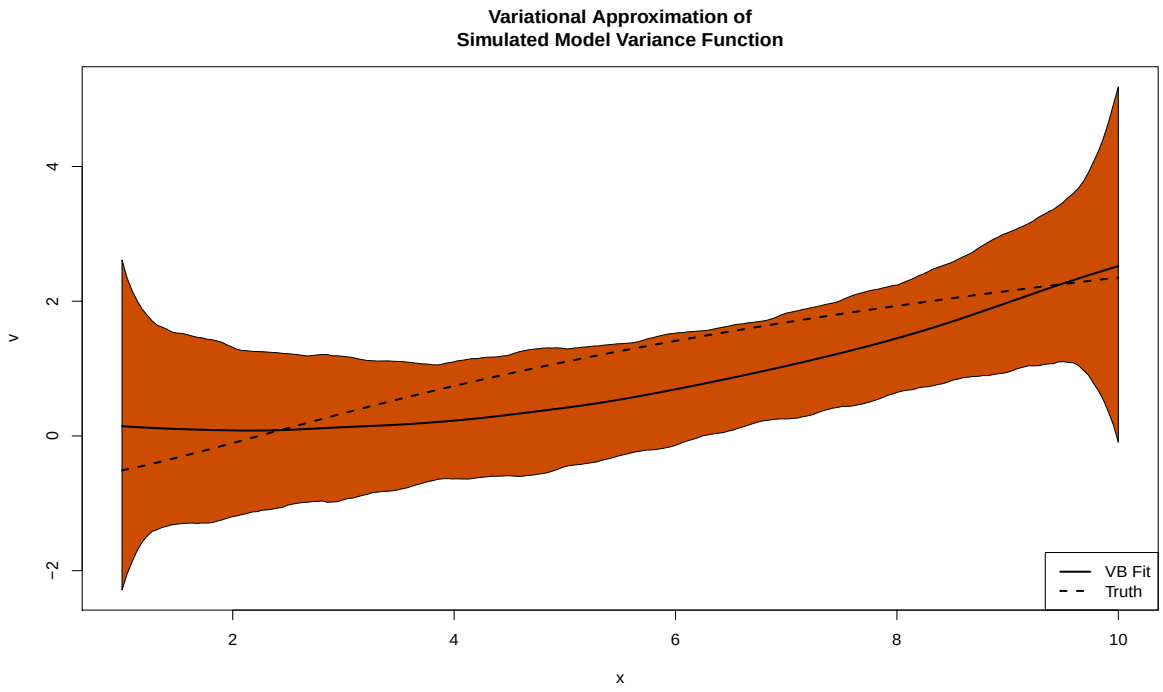


Figure 4.15: Approximate fit of the log model variance function when the radial variance function is $w_3(x, v)$. The shaded region represents a 95% credible region using the converged variational parameters of $\tilde{q}(\boldsymbol{\theta}_V)$

would need to be calculated. When $q(\boldsymbol{\theta}_V)$ is assumed to be Gaussian, this is equivalent to calculating the first $2p$ moments of a truncated Normal random variable, which can be directly computed. More complicated basis functions or interaction models could potentially lead to expectations that are not analytically available, making a variational approximation difficult. Future work will focus on implementing a variational approximation for a more general class of models that allows for both nonparametric representations based on the log model variance as well as interaction effects.

4.5 Variational Approximation for Spatially Adaptive Semiparametric Regression

We now detour from examining variational approximations for the models described in Chapter 2 in order to describe a novel result pertaining to the problem of semiparametric regression with a spatially adaptive penalty term. Our work in designing a variational Bayes approach to heteroskedastic semiparametric regression (Section 4.3) actually lends itself directly to this problem.

For standard semiparametric regression based on a mixed model framework, penalization of the spline basis functions is controlled by a global penalty term based on the ratio of random effect variance to error variance (Ruppert et al., 2003, chap. 4). However, the appropriateness of a global penalty approach can be called into question when the data suffers from rapid changes in curvature, heteroskedasticty, or a combination of the two. The model described in (13) can be thought of as having a form of spatially adaptivity since the error variance, σ^2, is allowed to be a smooth function of a covariate x. Depending on the model, this form of spatial adaptivity tends to affect the uncertainty bounds more than underlying fit of the mean function. This is somewhat expected given knowledge of the properties of mean estimates for fixed effect models under heteroskedasticity, namely that they are still unbiased (Barreto and Howland, 2005, chap. 19). However, a different spatially adaptive penalty structure is needed for cases where errors appear to be constant but there are dramatic changes in the underlying curvature of the model.

Consider the model set forth by Baladandayuthapani et al. (2005) for spatially adaptive penalized spline regression, restricted in this case to a single curve relationship:

$$y_i = \mathbf{X}_i^T \beta + \mathbf{Z}_i^T \mathbf{b} + \epsilon_\mathbf{i} = \mathbf{C}_i^T \boldsymbol{\theta} + \epsilon_\mathbf{i}$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\begin{aligned}
\mathbf{b} &\sim \mathcal{N}\left(0, \sigma^2_{b_k}\right) \;\; \forall k = 1, \ldots, K \\
\log(\sigma^2_{b_k}) &\sim \mathbf{X}^T_{B_k}\eta + \mathbf{Z}^T_{B_k}\mathbf{d} = \mathbf{C}^T_{B_k}\boldsymbol{\theta}_B. \\
\mathbf{d} &\sim \mathcal{N}\left(0, \sigma^2_d\right)
\end{aligned} \tag{109}$$

Here we assume a homoskedastic variance term, σ^2, but allow for the random effect variance σ^2_b to vary smoothly. Baladandayuthapani et al. (2005) show that models of this form are better able to handle oscillatory behavior and curvature changes than models with a global penalty. Examples of such data where a model like this would be appropriate can be found in Figure 4.16.

While the variance structure of this model is different from the heteroskedastic model in (13), the posterior conditional structure for the model parameters is strikingly similar. Assume priors of the form:

$$\begin{aligned}
\beta &\sim \mathcal{N}\left(0, \sigma^2_\beta \mathcal{I}_{p+1}\right) \\
\eta &\sim \mathcal{N}\left(0, \sigma^2_\eta \mathcal{I}_{r+1}\right) \\
\sigma^2 &\sim \mathcal{IG}\left(A, B\right) \\
\sigma^2_d &\sim \mathcal{IG}\left(A_d, B_d\right).
\end{aligned} \tag{110}$$

Let $p+1$ and $r+1$ denote the number of columns of $\mathbf{X}$ and $\mathbf{X}_B$. Here K and K_B are the column dimension of the random effect matrices $\mathbf{Z}$ and $\mathbf{Z}_B$. For illustrative purposes, consider our models to consist of two penalized spline fits of the form:

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \sum_{k=1}^{K} b_k \left(x_i - \kappa_k\right)^p_+ + \epsilon_i \\
\log(\sigma^2_{b_k}) &= \eta_0 + \eta_1 x_i + \cdots + \eta_r x_i^r + \sum_{l=1}^{K_B} d_l \left(x_i - \kappa^B_l\right)^r_+ .
\end{aligned} \tag{111}$$

Under these priors, the posterior conditionals for $\boldsymbol{\theta}$, $\boldsymbol{\theta}_B$, σ^2, and σ_d^2 are

$$\begin{aligned}
\boldsymbol{\theta} \mid \cdot &\sim \mathcal{N}\left(\mathbf{M}\mathbf{C}^T\mathbf{y}, \mathbf{M}\right) \\
\mathbf{M} &= \left(\frac{1}{\sigma^2}\mathbf{C}^T\mathbf{C} + \text{blockdiag}\left(\frac{1}{\sigma_\beta^2}\mathcal{I}_{p+1}, D(\sigma_{b_k}^2)^{-1}\right)\right)^{-1} \\
D(\sigma_{b_k}^2)^{-1} &= \text{diag}\left(\frac{1}{\sigma_{b_1}^2}, \frac{1}{\sigma_{b_2}^2}, \ldots, \frac{1}{\sigma_{b_K}^2}\right) \\
\sigma^2 \mid \cdot &\sim \mathcal{IG}\left(A + \frac{N}{2}, B + \frac{1}{2}\left\|\mathbf{y} - \mathbf{C}\boldsymbol{\theta}\right\|^2\right) \\
\sigma_d^2 \mid \cdot &\sim \mathcal{IG}\left(A_d + \frac{K_B}{2}, B_d + \frac{1}{2}\left\|\mathbf{d}\right\|^2\right) \\
p(\boldsymbol{\theta}_B \mid \cdot) &\propto \exp\left(-\frac{1}{2}\Big(\sum_{k=1}^{K}\mathbf{C}_{B_k}^T\boldsymbol{\theta}_B + \sum_{k=1}^{K} b_k \exp(-\mathbf{C}_{B_k}^T\boldsymbol{\theta}_B)\right. \\
&\qquad \left.+ {\boldsymbol{\theta}_B}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}_B}^{-1}\boldsymbol{\theta}_B\Big)\right). \qquad (112)
\end{aligned}$$

As with the posterior conditionals in (29), we have near-full conjugate structure with the exception of the posterior conditional associated with the second level of the semiparametric regression model ($\boldsymbol{\theta}_B$ in this case). However, the form of $p(\boldsymbol{\theta}_B \mid \cdot)$ is structurally the same as $p(\boldsymbol{\theta}_V \mid \cdot)$ from (29). This suggests that we can use the same variational approximation techniques described in Section 4.3 to construct an appropriate variational update algorithm for (110).

4.5.1 Variational Bayes Approximation

Let $\boldsymbol{\psi} = \left(\boldsymbol{\theta}^T, {\boldsymbol{\theta}_B}^T, \sigma^2, \sigma_d^2\right)^T$ be the vector of parameters of interest. Assume a product density restriction, described in (56), such that

$$q(\boldsymbol{\psi}) = q_1(\boldsymbol{\theta})q_2(\boldsymbol{\theta}_B)q_3(\sigma^2)q_4(\sigma_d^2). \qquad (113)$$

As before, we will drop the numerical subscript describing these variational densities for notational simplicity ($q_1(\boldsymbol{\theta}) = q(\boldsymbol{\theta})$). Using the relationship from (57), the optimal variational

densities are

$$\begin{aligned}
\boldsymbol{\theta} &\overset{q^*}{\sim} \mathcal{N}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\mathbf{C}^T\mathbf{y}, \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\right) \\
\sigma^2 &\overset{q^*}{\sim} \mathcal{IG}\left(A + \frac{N}{2}, B + \frac{1}{2}\left(\left\|\mathbf{y} - \mathbf{C}\mu_{q(\boldsymbol{\theta})}\right\|^2 + \text{trace}\left(\mathbf{C}^T\mathbf{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\right)\right)\right) \\
\sigma_d^2 &\overset{q^*}{\sim} \mathcal{IG}\left(A_d + \frac{K_B}{2}, B_d + \frac{1}{2}\left(\left\|\mu_{q(\mathbf{d})}\right\|^2 + \text{trace}\left(\boldsymbol{\Sigma}_{q(\mathbf{d})}\right)\right)\right) \\
q^*(\boldsymbol{\theta}_B) &\propto \exp\Bigg(-\frac{1}{2}\Big(\sum_{k=1}^{K} \mathbf{C}_{B_k}^T\boldsymbol{\theta}_B + \sum_{k=1}^{K}\left(\mu_{q(b_k)}^2 + \sigma_{q(b_k)}^2\right)\exp\left(-\mathbf{C}_{B_k}^T\boldsymbol{\theta}_B\right) \\
&\qquad + {\boldsymbol{\theta}_B}^T \text{blockdiag}\left(\frac{1}{\sigma_\eta^2}\mathcal{I}_{r+1}, \mu_{q(1/\sigma_d^2)}\mathcal{I}_{K_B}\right)\boldsymbol{\theta}_B\Big)\Bigg). \qquad (114)
\end{aligned}$$

Here, the variational covariance matrix associated with $q(\boldsymbol{\theta})$ is

$$\begin{aligned}
\Sigma_{q(\boldsymbol{\theta})} &= \left(\mu_{q(1/\sigma^2)}\mathbf{C}^T\mathbf{C} + \text{blockdiag}\left(\frac{1}{\sigma_\beta^2}\mathcal{I}_{p+1}, \boldsymbol{\Delta}\right)\right)^{-1} \\
\boldsymbol{\Delta} &= \text{diag}\left(E_{-\boldsymbol{\theta}}[\exp(-\mathbf{C}_{B_1}^T\boldsymbol{\theta}_B)], \ldots, E_{-\boldsymbol{\theta}}[\exp(-\mathbf{C}_{B_K}^T\boldsymbol{\theta}_B)]\right).
\end{aligned}$$

Denote

$$\begin{aligned}
B_{q(\sigma^2)} &= B + \frac{1}{2}\left(\left\|\mathbf{y} - \mathbf{C}\mu_{q(\boldsymbol{\theta})}\right\|^2 + \text{trace}\left(\mathbf{C}^T\mathbf{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\right)\right) \\
B_{q(\sigma_d^2)} &= B_d + \frac{1}{2}\left(\left\|\mu_{q(\mathbf{d})}\right\|^2 + \text{trace}\left(\boldsymbol{\Sigma}_{q(\mathbf{d})}\right)\right). \qquad (115)
\end{aligned}$$

In order to deal with the non-conjugate structure associated with $\boldsymbol{\theta}_B$, we use the same Laplace approximation technique described in Section 4.3.1. Define $h(\boldsymbol{\theta}_B \mid \cdot)$ as

$$\begin{aligned}
h(\boldsymbol{\theta}_B \mid \cdot) = \frac{1}{2}\Bigg(&\sum_{k=1}^{K} \mathbf{C}_{B_k}^T\boldsymbol{\theta}_B + \sum_{k=1}^{K}\left(\mu_{q(b_k)}^2 + \sigma_{q(b_k)}^2\right)\exp\left(-\mathbf{C}_{B_k}^T\boldsymbol{\theta}_B\right) \\
&+ {\boldsymbol{\theta}_B}^T \text{blockdiag}\left(\frac{1}{\sigma_\eta^2}\mathcal{I}_{r+1}, \mu_{q(1/\sigma_d^2)}\mathcal{I}_{K_B}\right)\boldsymbol{\theta}_B\Bigg). \qquad (116)
\end{aligned}$$

Let $\alpha = \text{argmin}\ h(\boldsymbol{\theta}_B \mid \cdot)$. It follows that the variational distribution of $\boldsymbol{\theta}_B$ can be approximated by a multivariate Gaussian distribution of the form

$$\boldsymbol{\theta}_B \overset{\tilde{q}}{\sim} \mathcal{N}\left(\alpha, \left(\frac{\mathbf{H}_{\boldsymbol{\theta}_\mathbf{B}}(\alpha)}{2}\right)^{-1}\right), \tag{117}$$

where $\mathbf{H}_{\boldsymbol{\theta}_\mathbf{V}}(\alpha)$ corresponds to the matrix of partial second derivatives of $h(\boldsymbol{\theta}_B \mid \cdot)$ evaluated at α. Under this approximate Gaussian form, the diagonal elements of the matrix $\boldsymbol{\Delta}$, denoted δ_k can be expressed as

$$\delta_k = \exp\left(-\mathbf{C}_{B_k}^T \mu_{q(\boldsymbol{\theta}_B)} + \frac{1}{2}\mathbf{C}_{B_k}^T \Sigma_{q(\boldsymbol{\theta}_B)} \mathbf{C}_{B_k}\right). \tag{118}$$

Using the variational densities described in (114) and the approximate form from (117) we construct the iterative method approximating spatially adaptive semiparametric regression found in Algorithm 5. As with the other algorithms presented here, convergence is assessed by monitoring the K-L divergence, $\log \underline{p}(\mathbf{y}; q)$, as defined by (55) as well as relative change in $\|\mu_{q(\boldsymbol{\theta})}\|^2$ and $\|\mu_{q(\boldsymbol{\theta}_B)}\|^2$. The explicit formula for the K-L divergence of this model is

$$\begin{aligned}
\log \underline{p}(\mathbf{y}; q) &= \int_\Psi q(\boldsymbol{\psi}) \left(\log(p(\mathbf{y} \mid \boldsymbol{\psi}) p(\boldsymbol{\psi})) - \log(q(\boldsymbol{\psi}))\right) d\boldsymbol{\psi} \\
&= -\frac{N}{2}\log(2\pi) - \frac{p+1}{2}\log(\sigma_\beta^2) - \frac{r+1}{2}\log(\sigma_\eta^2) \\
&\quad - \frac{1}{2}\mu_{q(1/\sigma^2)}\left(\|\mathbf{y} - \mathbf{C}\mu_{q(\boldsymbol{\theta})}\|^2 + \text{trace}\left(\mathbf{C}^T\mathbf{C}\boldsymbol{\Sigma}_{q(\theta)}\right)\right) \\
&\quad - \frac{1}{2}\sum_{k=1}^K \mathbf{C}_{B_k}^T \mu_{q(\boldsymbol{\theta}_B)} - \frac{1}{2\sigma_\beta^2}\left(\|\mu_{q(\beta)}\|^2 + \text{trace}(\boldsymbol{\Sigma}_{q(\beta)})\right) \\
&\quad - \frac{1}{2}\sum_{k=1}^K \left(\mu_{q(b_k)}^2 + \sigma_{q(b_k)}^2\right) \exp\left(-\mathbf{C}_{B_k}^T \mu_{q(\boldsymbol{\theta}_B)} + \frac{1}{2}\mathbf{C}_{B_k}^T \boldsymbol{\Sigma}_{q(\boldsymbol{\theta}_B)} \mathbf{C}_{B_k}\right) \\
&\quad - \frac{1}{2\sigma_\eta^2}\left(\|\mu_{q(\eta)}\|^2 + \text{trace}(\boldsymbol{\Sigma}_{q(\eta)})\right) - \frac{1}{2}\mu_{q(1/\sigma_d^2)}\left(\|\mu_{q(\mathbf{d})}\|^2 + \text{trace}(\boldsymbol{\Sigma}_{q(\mathbf{d})})\right) \\
&\quad + A_d \log(B_d) - \log(\Gamma(A_d)) - \mu_{q(1/\sigma_d^2)} B_d + \frac{p+K+1}{2} \\
&\quad + A\log(B) - \log(\Gamma(A)) - \mu_{q(1/\sigma^2)} B + \frac{r+K_B+1}{2}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{2}\log\left(\left|\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\right|\right) + \frac{1}{2}\log\left(\left|\boldsymbol{\Sigma}_{q(\boldsymbol{\theta}_B)}\right|\right) \\
&- (A + \frac{N}{2})\log(B_{q(\sigma^2)}) + \log\left(\Gamma\left(A + \frac{N}{2}\right)\right) + B_{q(\sigma^2)}\mu_{q(1/\sigma^2)} \\
&- (A_d + \frac{K_B}{2})\log(B_{q(\sigma_d^2)}) + \log\left(\Gamma\left(A_d + \frac{K_B}{2}\right)\right) + B_{q(\sigma_d^2)}\mu_{q(1/\sigma_d^2)}.
\end{aligned}
$$

Algorithm 5 Iterative method for determining the optimal variational distributions for the parameters of a spatially adaptive semiparametric regression via penalized splines model

Initialize: $\mu_{q(\boldsymbol{\theta})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$, $B_{q(\sigma_d^2)}$,$B_{q(\sigma^2)}$

repeat

  Compute $\alpha = \text{argmin } h(\boldsymbol{\theta}_B \mid \cdot)$, $\mathbf{H}_{\boldsymbol{\theta}_B}(\alpha)$

  $\mu_{\tilde{q}(\boldsymbol{\theta}_B)} \leftarrow \alpha$

  $\boldsymbol{\Sigma}_{\tilde{q}(\boldsymbol{\theta}_B)} \leftarrow \left(\frac{\mathbf{H}_{\boldsymbol{\theta}_V}(\alpha)}{2}\right)^{-1}$

  $B_{q(\sigma_d^2)} \leftarrow B_d + \frac{1}{2}\|\mu_{q(\mathbf{d})}\|^2 + \frac{1}{2}\text{trace}(\boldsymbol{\Sigma}_{q(\mathbf{d})})$

  $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \leftarrow \left(\mu_{q(1/\sigma^2)}\mathbf{C}^T\mathbf{C} + \text{blockdiag}\left(\frac{1}{\sigma_\beta^2}\mathcal{I}_{p+1}, \boldsymbol{\Delta}\right)\right)^{-1}$

  $\mu_{q(\boldsymbol{\theta})} \leftarrow \Sigma_{q(\boldsymbol{\theta})}\mathbf{C}^T\mathbf{y}$

  $B_{q(\sigma^2)} \leftarrow B + \frac{1}{2}\left(\left\|\mathbf{y} - \mathbf{C}\mu_{q(\boldsymbol{\theta})}\right\|^2 + \text{trace}\left(\mathbf{C}^T\mathbf{C}\Sigma_{q(\boldsymbol{\theta})}\right)\right)$

until Convergence is Reached

Construct parameter estimates using mean of variational distributions

4.5.2 Spatially Adaptive Example

To illustrate the method described in Algorithm 5, we consider a simulated data example akin to the one presented in Baladandayuthapani et al. (2005). Consider a true mean function

$$
m_j(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi(1+2^{(9-4j)/5})}{x+2^{(9-4j)/5}}\right), \tag{119}
$$

over the region $x \in [0,1]$ where j controls spatial variability and oscillatory behavior of the function. Higher values of j will lead to increased oscillation towards 0. Simulated datasets

of the form $y_i = m_j(x_i) + \epsilon_i$ are created for both a low spatial variability ($j = 3$) and a high spatial variability ($j = 6$). Each example consists of 800 data points evaluated at equally spaced values of x across $[0, 1]$. The error term ϵ is distributed $\mathcal{N}(0, 0.04)$.

Figure 4.16 contains the simulated data for both $j = 3$ and $j = 6$ cases. A semiparametric, global penalty model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+^2 + \epsilon_i, \tag{120}$$

is fit via the variational approximation algorithm described in Algorithm 2. Equally spaced knots from 0.01 to 0.99 are used with $K = 30$ and $K = 90$ for the low-variability and the high-variability cases ($j = 3$ and $j = 6$) respectively. Oscillatory data of this type highlight the difficulty that global penalty models can have with rapid changes in curvature. For the low-variability case, the global penalty model appears to fit relatively well. For the high-variability case, the global penalty has great difficulty fitting the data at the low regions of x where the oscillation is heaviest. Also, the fit associated with the higher region of x appears to suffer from under-smoothing given the local behavior.

Figure 4.17 shows the corresponding variational approximations of the spatially adaptive semiparametric model fits for the simulated data presented here. In addition to the model above, the penalized spline model associated with the σ^2 term is also fit using a truncated quadratic spline basis with $K_B = 5$ and $K_B = 15$ for the low and high variability cases respectably. For the low-variability case, the spatially adaptive model is comparable to the global penalty model, as one would expect. In the high variability case, the spatially adaptive variational approximation does a considerably better job at dealing with the high oscillation region near $x = 0$. Granted, there are limits to how well penalized splines can fit a rapidly changing structure like this due to knot choices and sample size, but the credible regions associated with this approximation appears to provide adequate coverage throughout this

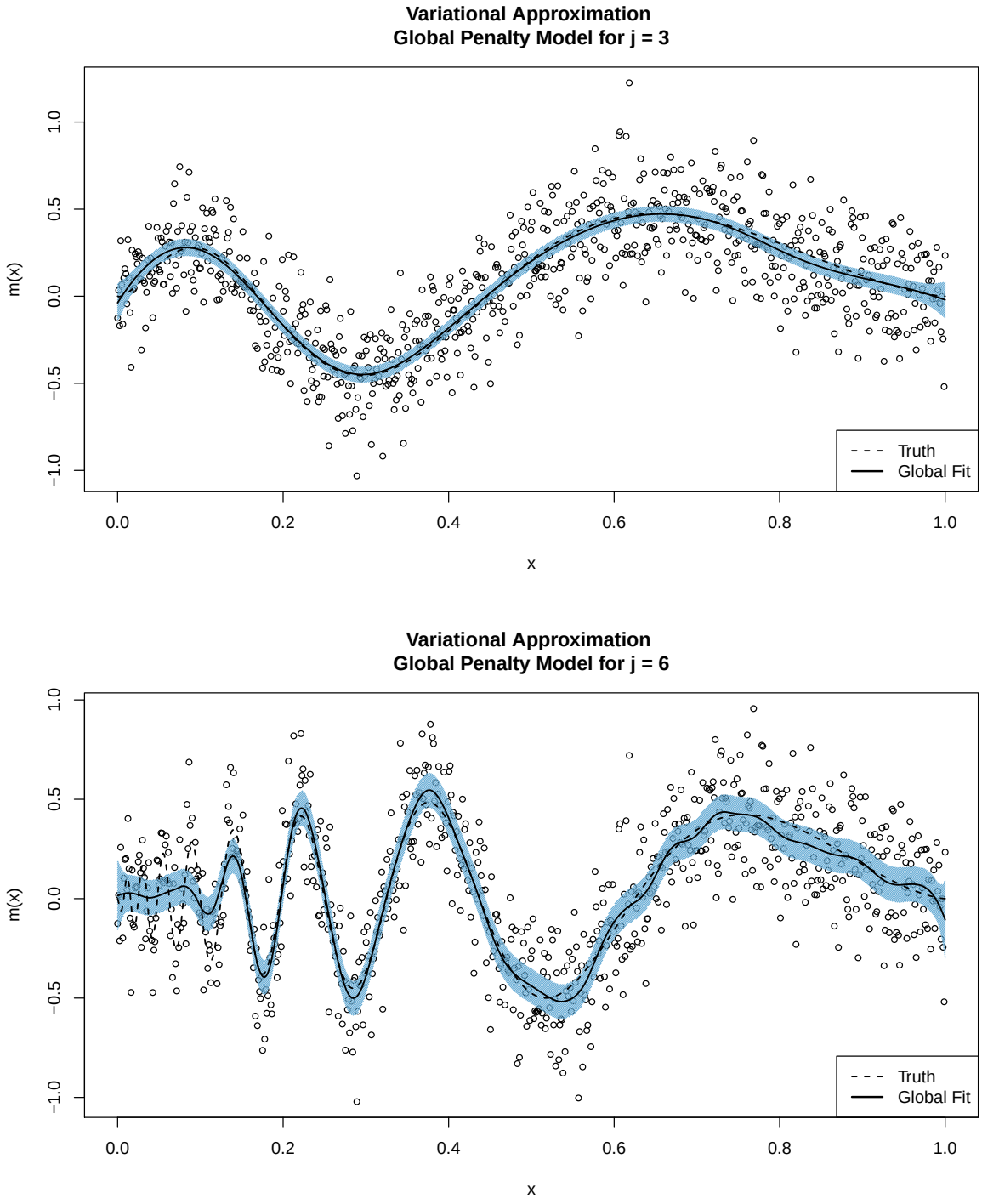


Figure 4.16: Variational approximations for a standard penalized spline model using a global penalty term fit via a mixed model framework for low and high oscillation scenarios. The top figure corresponds to data generated from $m_3(x)$ while the bottom corresponds to $m_6(x)$. The fit for $j = 3$ used $K = 30$ equally spaced knots while $j = 6$ used $K = 90$.

example. Also, the fit associated with the smoother portions of the data better reflect the local behavior of the underlying mean function and do not appear to suffer from the same under smoothing present in the global version. These finding are consistent with the more exact implementations found in Baladandayuthapani et al. (2005).

4.5.3 Extensions

The work presented here follows in the footsteps of Baladandayuthapani et al. (2005) and only considers a spatially adaptive model with a single penalty term. This corresponds to adding an additional level of the model to describe the smooth behavior of $\log(\sigma_b^2)$. One of the advantages of the mixed model representation of penalized splines is that multiple non-parametric components and thus multiple penalty terms are allowed in the model. These are controlled through the addition of new random effects with a separate variance parameter. The spatially adaptive framework discussed here can easily be extended to consider multiple penalties, some of which may be modeled as spatially varying. Operating under the assumption that the distinct random effect variances are independent, this would result in multiple $\log(\sigma_{b_l}^2)$ structures. Each of theses levels would contain their own model parameters $\boldsymbol{\theta}_{Bl}$ which may include additional random effect variances (akin to the σ_d^2 term presented previously). Deriving the posterior conditional structures of this extended model would be relatively straightforward using standard methods.

Each of the $\boldsymbol{\theta}_{Bl}$ would result in a non-conjugate posterior distribution, $\boldsymbol{\theta}_{Bl} \mid \cdot$. Using the variational approximation method presented here, these would be handled through an additional step of an embedded Laplace approximation. Care needs to be exercised here since the effect of the inclusion of multiple Laplace approximations of this form on the convergence and computational properties of the variational approximation has not been studied. This is an open question that will be investigated further in future work.

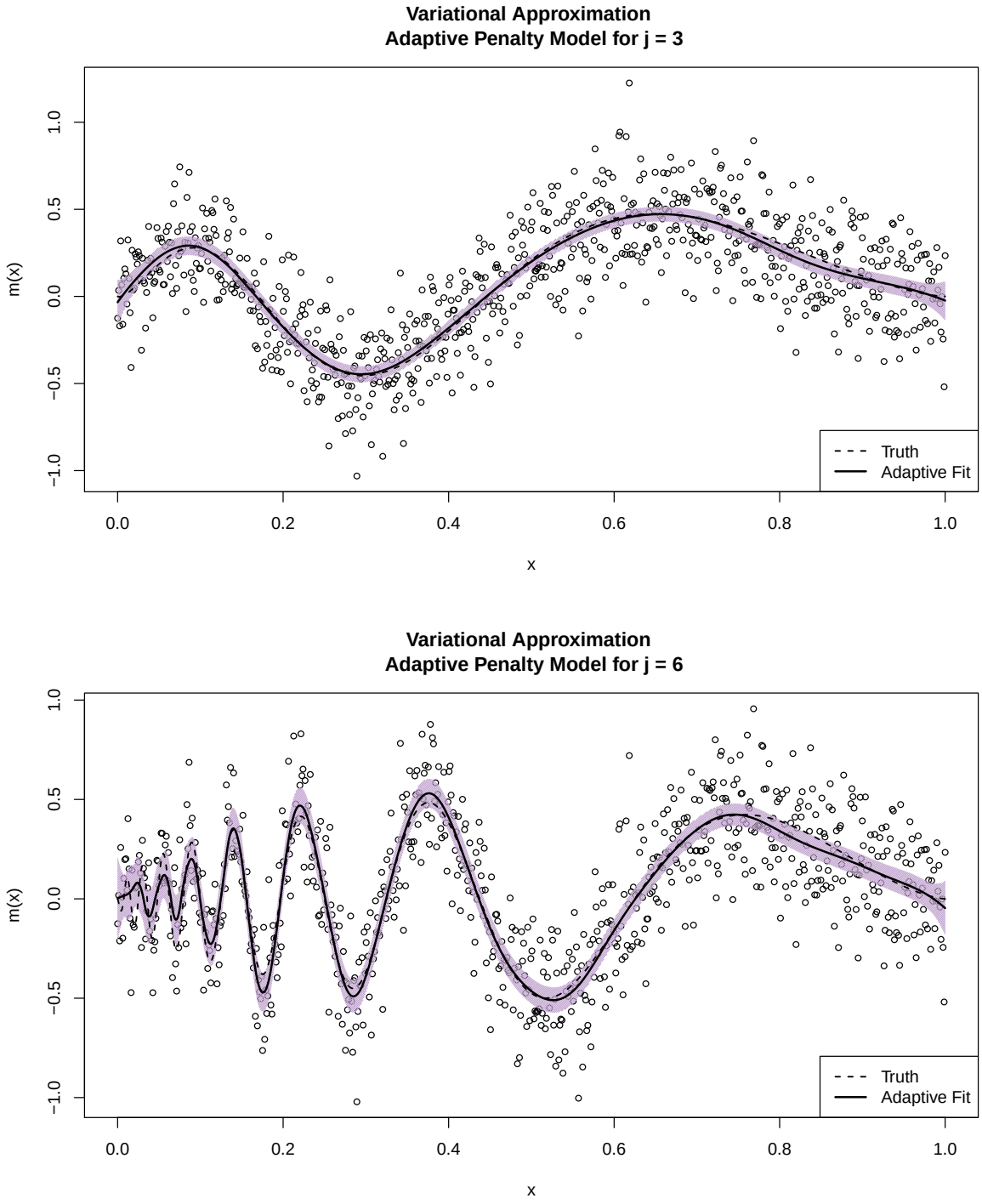


Figure 4.17: Variational approximations for a penalized spline model using a spatially adaptive penalty term fit via a mixed model framework for low and high oscillation scenarios. The top figure corresponds to data generated from $m_3(x)$ while the bottom corresponds to $m_6(x)$. The fit for $j = 3$ used $K = 30$ and $K_B = 5$ while $j = 6$ used $K = 90$ and $K_B = 15$.

A natural extension of the work presented here and in Section 4.3 would be to construct a variational method for approximating spatially adaptive penalized spline models with heteroskedastic errors as described in Crainiceanu et al. (2007). These models describe the smooth structure of both $\log(\sigma^2)$ and $\log(\sigma_b^2)$. This requires the inclusion of model parameters $\boldsymbol{\theta}_V$ and $\boldsymbol{\theta}_B$ that induce non-conjugate posterior conditional forms, $\boldsymbol{\theta}_V \mid \cdot$ and $\boldsymbol{\theta}_B \mid \cdot$. Our variational methods seem appropriate for constructing a fast approximation of this model. However, care needs to be taken since this would require at least two embedded Laplace approximations and careful study of the effects on the convergence and computational properties of the approximation would need to be conducted.

CHAPTER 5

CASE STUDY OF EXPERIMENTAL SAXS DATA

We showcase the various methodologies presented here in the context of SAXS experimental data. SAXS presents a variety of interesting statistical problems. One example is investigating the correlation structure of pixels on the two-dimensional sensor (Breidt et al., 2012). The motivating goal of the work presented here has been to create well-documented frameworks that can be used to address two classes of inferential questions. The first class is concerned with "data quality". These questions typically center on the effect of experimental covariates such as concentration and exposure time on the SAXS data. The second class concerns inference on actual physical characteristics of the molecule of interest. This chapter will provide multiple examples to illustrate the use of our work for real SAXS problems. This is by no means comprehensive but rather should be considered a primer on analyzing SAXS data.

5.1 Inference on Experimental Factors

Experimental factors often have unforseen consequences to the quality of SAXS data. Since the end goal is measurement of a directly unobservable quantity (the sample molecule's shape), degradation of the data quality can lead to improper inference about the true shape of the molecule. These issues can include improper calibration, sensor failure, overexposure, particle aggregation, and radiation damage. In the presence of multiple exposures under known experimental conditions, the frameworks presented here can be tailored to answer questions regarding these issues. For basic illustration, we present two common cases and discuss briefly how they can be addressed within the model frameworks presented in this

work. Full, in-depth analyses of specific SAXS data are omitted for now as the goal of this section is to illustrate rather than investigate.

5.1.1 Detection of Concentration Effects

SAXS experiments typically consist of exposures taken across multiple concentrations. This is typically done in an effort to reduce noise in the measurements, particularly at high angles. Ideally, the shape inference about the molecule of interest should be the same regardless of concentration. Theoretically, concentration has an additive effect to the log intensity data (multiplicative on the original scale). However the complex macromolecules that are often the subject of these experiments do not always behave ideally under concentration changes. One possible problem is that increases in concentration can cause aggregation of particles, inherently changing the shape that the experiment "sees". Concentration changes can also potentially cause violations of the assumption that all molecules are tumbling freely in solution and thus interact with each other. Detection of this behavior is important to ensure valid inference is being drawn across concentrations.

Consider sampled versions of two 7-second exposures of a nucleosome assembly protein (NAP), displayed in Figure 5.1. The top curve (purple) is taken at 20 mg/ml and the bottom curve (green) is taken at 5 mg/ml. The dramatic smoothness of the data can be attributed to the radial smoothing and processing that occurs at the point of collection. To the naked eye, the curves only seem to differ by a constant shift with some potential deviation at low angles ($s \leq 0.05$). While there is slight evidence of heteroskedasticiy for the low concentration data, we consider a homoskedastic model here to ease illustration.

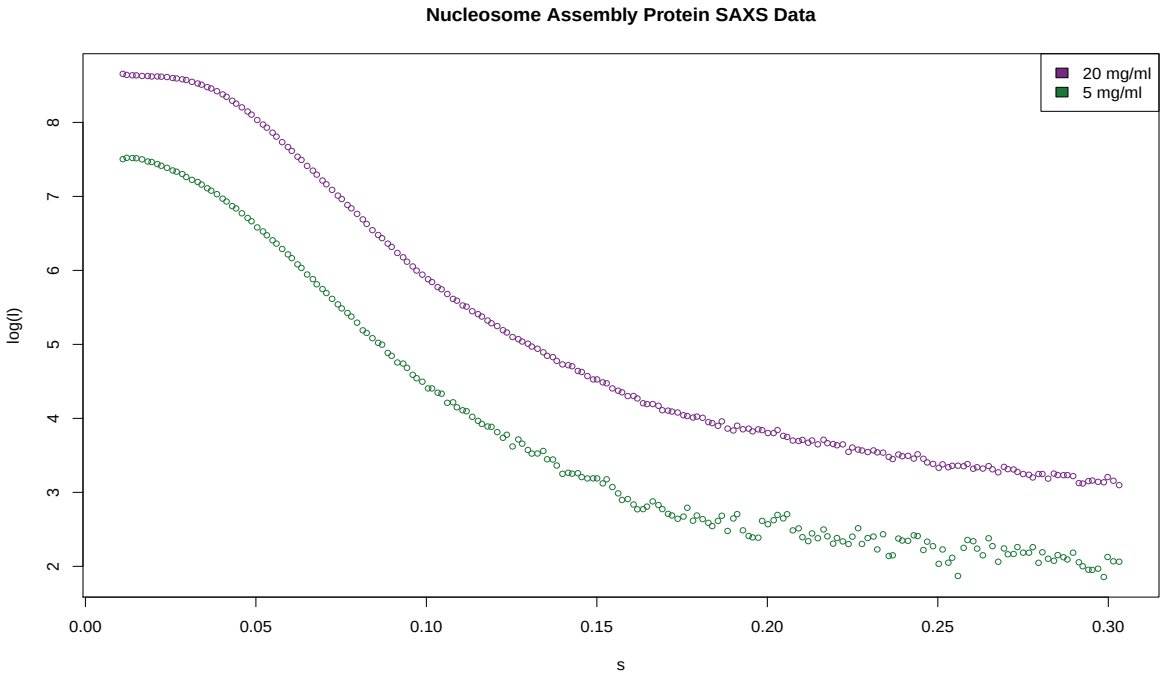


Figure 5.1: Data corresponding to two exposures of a sample of nucleosome assembly protein. The top curve was taken at 20 mg/ml and the bottom was taken at 5 mg/ml. Both data sets were exposed to high-intensity X-rays for 7-seconds.

Testing for non-standard concentration effects can be posed as a model selection problem. Consider the models

$$
\begin{aligned}
M_0: \;& y_i = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 \text{Conc}_i + \sum_{k=1}^{K} b_{1k} \left(s_i - \kappa_k\right)_+^2 + \epsilon_i \\
M_1: \;& y_i = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 \text{Conc}_i + \beta_4 \text{Conc}_i s_i + \beta_5 \text{Conc}_i s_i^2 \\
& \quad + \sum_{k=1}^{K} b_{1k} \left(s_i - \kappa_k\right)_+^2 + \epsilon_i \\
M_2: \;& y_i = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 \text{Conc}_i + \beta_4 \text{Conc}_i s_i + \beta_5 \text{Conc}_i s_i^2 \\
& \quad + \sum_{k=1}^{K} b_{1k} \left(s_i - \kappa_k\right)_+^2 + \sum_{k=1}^{K} b_{2k} \text{Conc}_i \left(s_i - \kappa_k\right)_+^2 + \epsilon_i. \qquad (121)
\end{aligned}
$$

Here $\{b_{1k}\}_{k=1}^K$ and $\{b_{2k}\}_{k=1}^K$ correspond to sets of random effects with common variance. The parameter Conc_i is the concentration of the ith measurement. Model M_0 can be considered the null model for the NAP case where the concentration only has an additive effect on log intensity. Models M_1 and M_2 represent expanded interaction models with the latter corresponding to a full parametric-by-nonparametric interaction.

Figures 5.2, 5.3, and 5.4 display the fits of each model to the NAP data. For each model, a 10000 iteration Gibbs sampler was used (Section 3.1) with a 1000 iteration burn-in period. Through graphical inspection alone, both the null model and the fixed effect interaction model are unable to fully fit the data, particularly given the smooth nature at small values of s. M_0's lack of fit for s in the regions $[0, 0.05]$ and $[0.20, 0.30]$ suggests a nonstandard concentration effect in the data. Figure 5.3 shows that M_1 is better able to fit the data for both concentrations at high values of s but still suffers from dramatic lack of fit near the origin. Figure 5.4 shows that the full parametric-by-nonparametric interaction model, M_2, is able to best fit the data for all values of s across both concentrations.

For this example, there is enough evidence to suggest the presence of a nonstandard concentration effect that may warrant scientific investigation. More rigorous model selection can be used to better answer this question. Under the Bayesian representations used in the paper, techniques such as Bayes factor or DIC can be used here to decide between the models in question (Gelman et al., 2013, chap. 4). For this example, the DIC for M_0 is -817.9219, the DIC for M_1 is -1065.45, and the DIC for M_2 is -1152.061. Lower values of DIC suggest stronger evidence for that model, implying that the fully parametric-by-nonparametric interaction model of M_2 is appropriate.

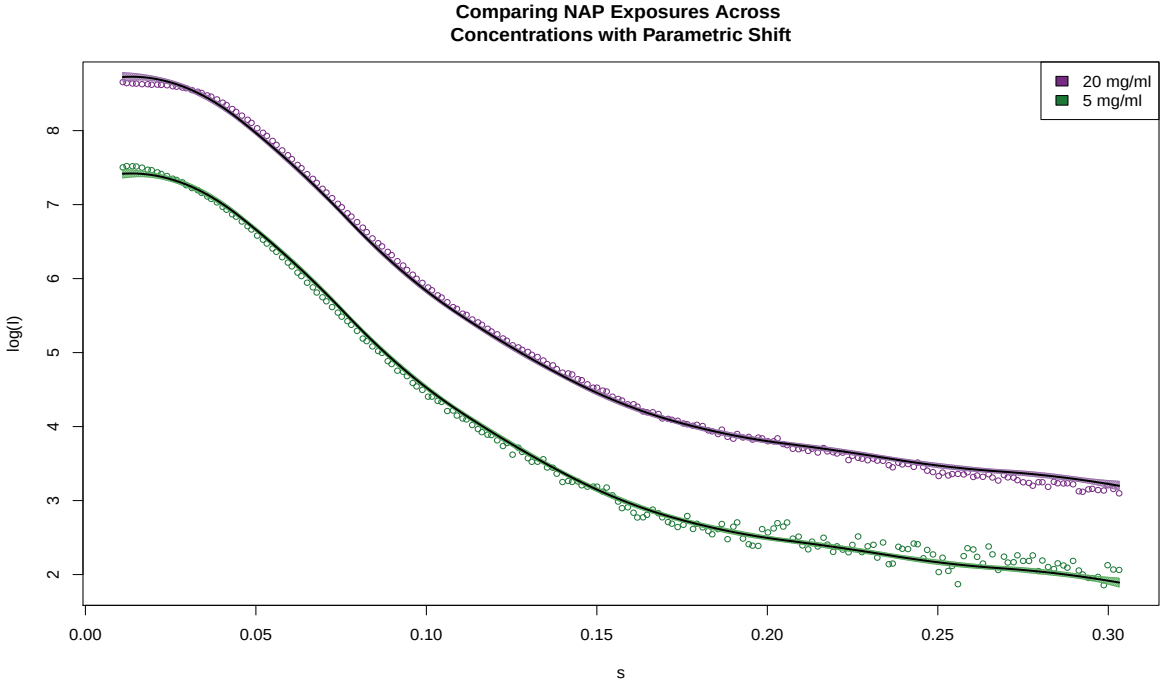


Figure 5.2: MCMC estimate via Gibbs sampling of the model M_0 for two exposures of a nucelosome assembly protein.

5.1.2 Detection of Exposure Effects

Another common experimental condition that carries strong influence over the quality of SAXS data is the amount of time each sample is bombarded by X-rays. The main concern is that repeated exposures, particularly long exposures, can cause damage to the sample

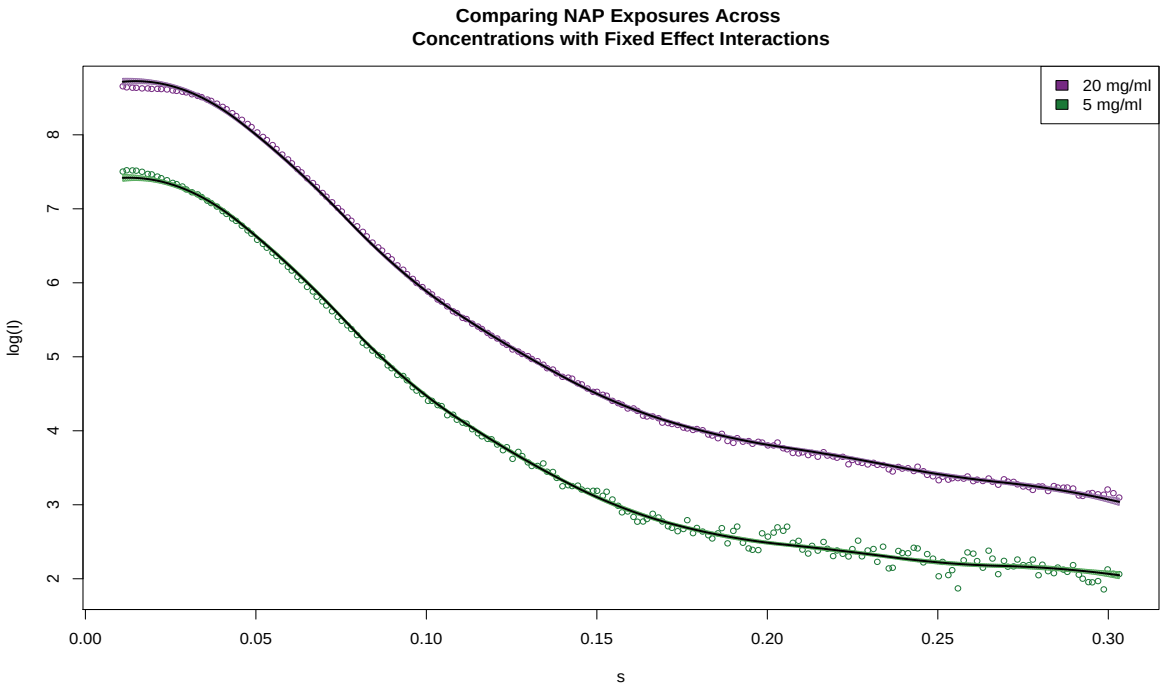


Figure 5.3: MCMC estimate via Gibbs sampling of the model M_1 for two exposures of a nucelosome assembly protein.

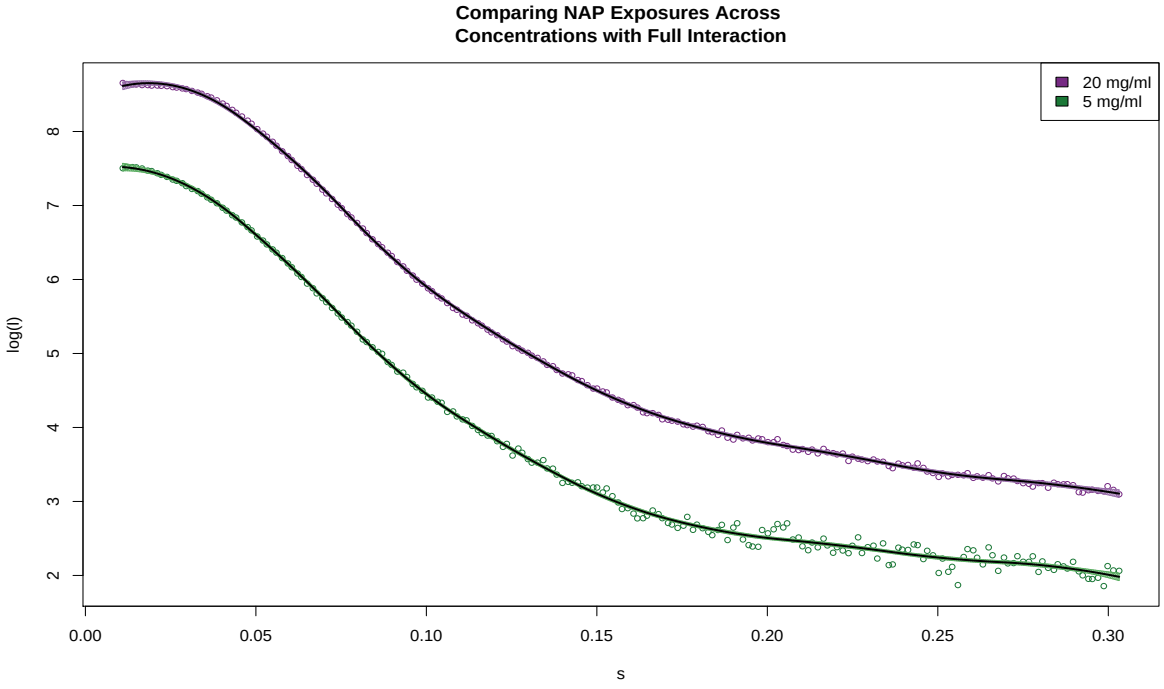


Figure 5.4: MCMC estimate via Gibbs sampling of the model M_2 for two exposures of a nucelosome assembly protein.

in question. This could change the molecular structure of interest, leading experimental scientists to make incorrect inference. As with the effect of concentration, testing for exposure time effects and radiation damage can be easily framed as a model selection problem under the frameworks presented here.

Consider a set of four SAXS exposures of the `H2AH2B` complex taken sequentially. The exposures are all taken at 11.8 mg/ml with exposure times of 7-seconds, 7-seconds, 70-seconds, and 7-seconds. The four exposures are plotted separately in Figure 5.5. The typical purpose of a long exposure is to obtain better "high angle information", referring to the variance dampening effect seen in the bottom left panel of Figure 5.5. This is done at the cost of erratic low angle behavior, which is often due to physical limitations of the pixels of the sensor plate associated with small s. Comparing SAXS exposures sequentially typically carries concerns for of radiation damage which are compounded following a long exposure. Figure 5.6 plots the 7-second exposures that directly precede and follow the 70-second case. Ideally, these data would line up perfectly since nothing has change except their order in the series. However, there appears to be structural difference of the two data sources for values of $s > 0.15$.

As with the concentration effect example of the preceding section, inferring the presence of a significant exposure effect can be viewed as a model selection problem. Consider the models

$$\begin{aligned} M_0 : y_i &= \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \sum_{k=1}^{K} b_k \left(s_i - \kappa_k\right)_+^2 + \epsilon \\ M_1 : y_i &= \beta_0 + \beta_1 s_i + \beta_2 s_i + \mathrm{I}_i^{\mathrm{Post}} \left(\beta_3 + \beta_4 s_i + \beta_5 s_i^2\right) \\ &\quad + \sum_{k=1}^{K} b_{1k} \left(s_i - \kappa_k\right)_+^2 + \sum_{k=1}^{K} b_{2k} \mathrm{I}_i^{\mathrm{Post}} \left(s_i - \kappa_k\right)_+^2 + \epsilon_i, \end{aligned} \tag{122}$$

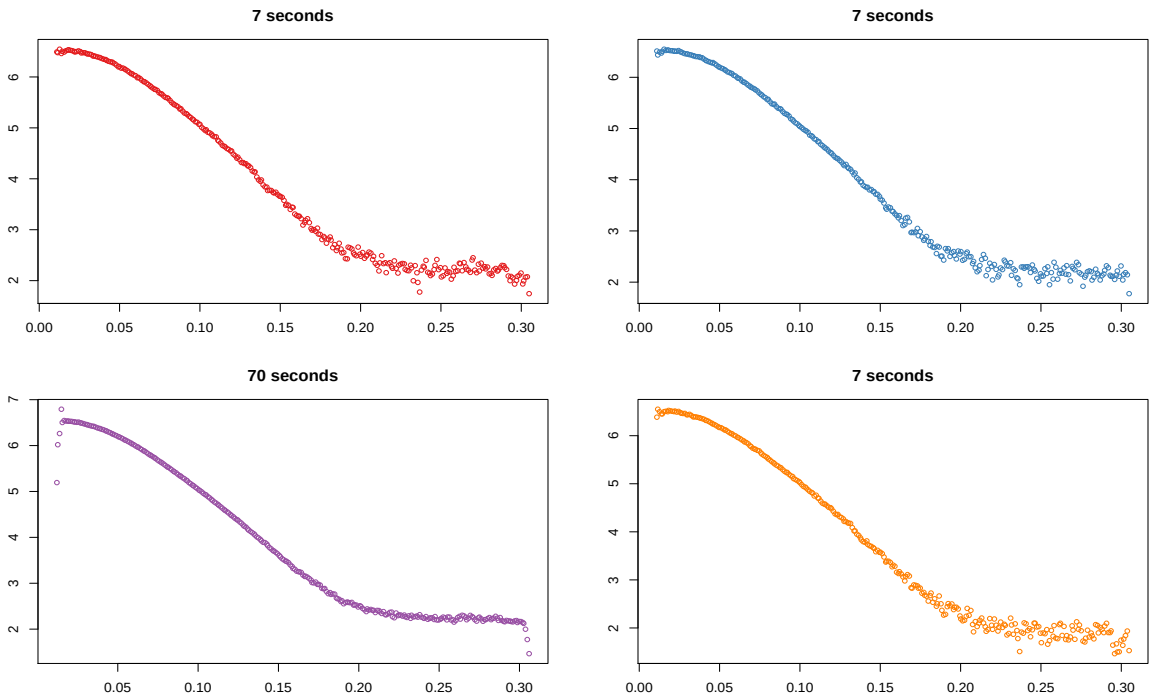


Figure 5.5: Data from a 7-7-70-7 exposure series of the H2AH2B complex. The exposures are taken sequentially moving from right to left, top to bottom.

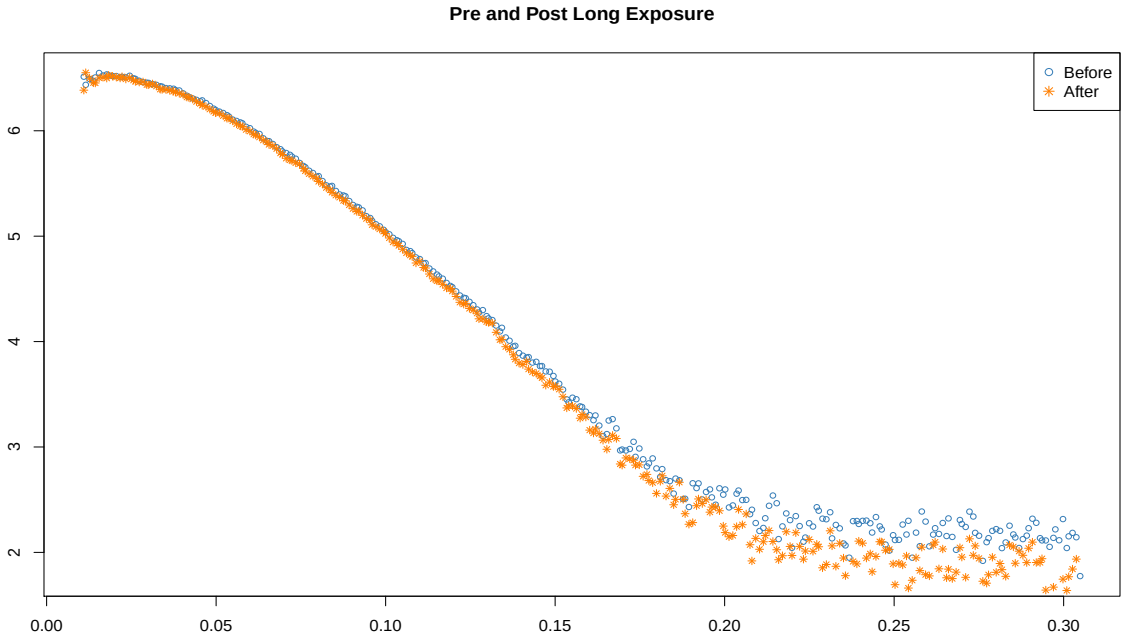


Figure 5.6: The SAXS data in this figure corresponds to 7-second exposures taken before the 70-second exposure (circle) and after (asterisk).

where $\mathrm{I}_i^{\mathrm{Post}}$ is an indicator function that takes values of 1 if the ith measurement is from the post long exposure data. Model M_1 is equivalent to fitting two separate functions to the data depending on which source they arose from. Model M_0 assumes there is no relevant information to be gained by knowing from which source each data point came. There are many models that could serve as alternatives to M_0 that are not as dramatic in their separation as M_1, possibly by allowing the exposure effect to only "kick in" for the large s regions. However M_0 and M_1 are more illustrative than practical in this example.

The MCMC procedure for handling heteroskedastic semiparametric regression (Section 3.2) is used for both M_0 and M_1. For both the mean and variance levels, $K = K_V = 10$ knots placed at equally spaced quantiles of $\{s_i\}_{i=1}^N$ ranging from 0.1 to 0.9 are used for a truncated quadratic polynomial basis. The procedures are run for 10000 iterations with a burn-in of 1000. Figure 5.7 contains the resulting fit for the null model M_0. As expected, the null model essentially splits the difference between pre- and post- long exposure data for large values of s. Figure 5.8 shows the fit for the alternative model M_1. Graphically, we see that M_1 fits both data sets quite well. This is to be expected since M_1 generically allows for separate nonparametric functions. Whether the gain in goodness-of-fit is worth the additional model complexity can be determined by more rigorous model selection methods (Bayes factor, DIC, etc.) as described in the concentration effect example. For this case, the DIC for M_0 is -1586.027 while the DIC for M_1 is -1853.532, supporting the exposure order interaction model.

5.2 Inference on Molecular Properties

The ultimate goal of SAXS experiments is to reconstruct structural shape information about complex macromolecules from one-dimensional intensity curves. This traditionally focuses on low-resolution information that describes general shape characteristics. Two such

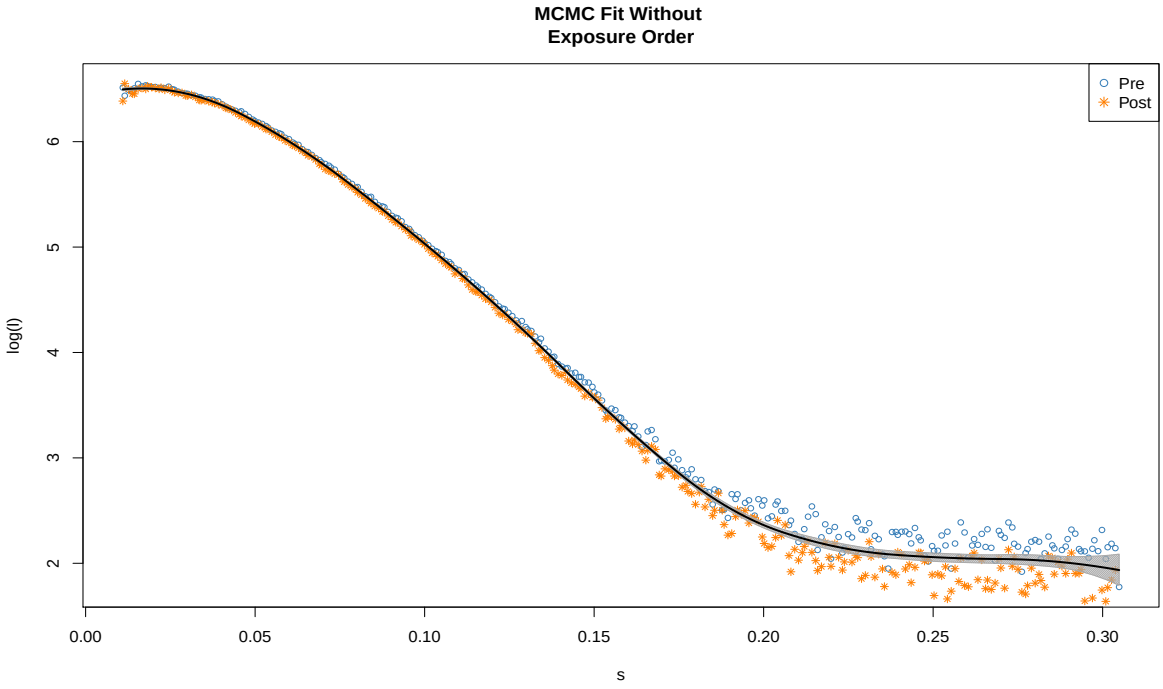


Figure 5.7: The null model M_0 is fit to the `H2AH2B` data. The heteroskedastic MCMC procedure was run for 10000 iterations with a burn-in of 1000.The shaded region corresponds to the 95% credible region of the fit of the mean level data.

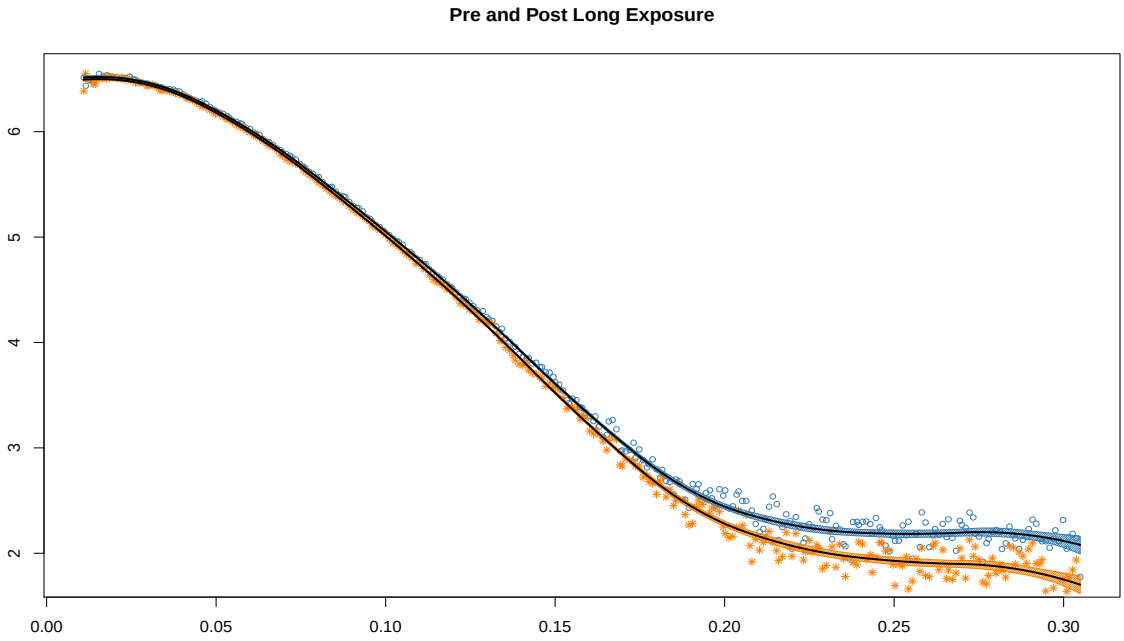


Figure 5.8: The null model M_1 is fit to the `H2AH2B` data. The heteroskedastic MCMC procedure was run for 10000 iterations with a burn-in of 1000.The shaded region corresponds to the 95% credible region of the fit of the mean level data.

physical parameters that we will focus on are $\log(I(0))$, the log intensity intercept and R_g, the radius of gyration.

The log intensity intercept, $\log(I(0))$, is of interest as a measure of a molecule's molecular weight. Given fixed concentration and experimental conditions, $\log(I(0))$ is proportional up to a known constant to molecular weight (Fischer et al., 2009). Including this parameter in any of the frameworks presented here is automatic. The regression models used to handle log intensity data include the standard intercept term β_0. Inference and uncertainty estimates about β_0 are obtainable through standard methods. Non-distance related effects, such as a concentration fixed effect, can be included. In this case, depending on the model specification, $\log(I(0))$ for a specific set of experimental conditions would be some linear combination of parameter estimates.

The radius of gyration, R_g, is defined as the square root of the ratio of the particle's moment of inertia and mass. Conceptually, the radius of gyration is representative of the radius of the time-averaged "shell" that would encompass the particle as it rotates freely. Standard Guinier analysis defines the squared radius of gyration over a small neighborhood near the origin as

$$I(s) = I(0) \exp\left(-\frac{1}{3} s^2 R_g^2\right) \tag{123}$$

where s is the measure of radial distance (angle in SAXS terminology) and $I(s)$ is the intensity function Guinier (1939). To estimate R_g^2 for a single SAXS exposure in our framework, one has two options. The first is to use a truncated quadratic spline basis while removing the s fixed effect. That is, the mean level model for log intensity data would be

$$\log(I)_i = \beta_0 + \beta_2 s_i^2 + \sum_{k=1}^{K} b_k (s_i - \kappa_k)_+^2. \tag{124}$$

While losing the flexibility of the linear function in the basis, there is theoretical justification for the locally quadratic structure $\beta_0 + \beta_2 s^2$ (Guinier, 1939). Alternatively, one could model $\log(I)$ vs s^2 and use a fully defined truncated linear spline basis.

For either case, estimation of R_g^2 becomes contingent on the span of the local neighborhood over which the Guinier relationship is appropriate. Using low-rank penalized splines, this is equivalent to selecting the location of the knot κ_1. In the work presented here, κ_1 is generally chosen to best fit the data rather than meeting any "optimal window" criteria in a scientific sense. Selecting the best window for estimating the radius gyration is currently being studied by Cody Alasker, a fellow PhD student at Colorado State University, with Jay Breidt and Mark van der Woerd.

5.2.1 `H2AH2B` Example

To illustrate the estimation of physical characteristics, we consider a sampled single exposure of the `H2AH2B` complex (Isenberg, 1979). This experiment consists of a 4 mg/ml sample of the `H2AH2B` complex irradiated for 7-seconds. This data displays pronounced heteroskedasticity for larger values of s. We present results from both the MCMC (Section 3.2) and variational approximation (Section 4.3) of a heterokedastic semiparametric regression model for this data. For comparison, we also present the homoskedastic versions as well (Sections 3.1 and 4.2). The Gibbs sampler for the constant errors model was run for 20000 iterations with a burn-in of 2000 for both models. The hybrid Gibbs-DRAM procedure was run for 40000 iterations with a burn-in of 4000.

For both error cases, we treat the mean-level responses under the model

$$y_i = \beta_0 + \beta_2 s_i^2 + \sum_{k=1}^{K} b_k \left(s_i - \kappa_k\right)_+^2 + \epsilon_i, \tag{125}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ or $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. For this example we use $K = 10$ knots at equally spaced quantiles of s ranging from 0.1 to 0.99. For the heteroskedastic version, the log model variance is modeled as

$$\log(\sigma_i^2) = \delta_0 + \delta_1 s_i + \delta_2 s_i^2 + \sum_{k=1}^{K_V} c_k \left(s_i - \kappa_{V_k}\right)_+^2 . \tag{126}$$

The same knots are used for the variance level as the mean level. All hyperparameter values are the same as previous examples.

		Constant Errors	**Heteroskedastic**
$\widehat{\beta_0}$	Estimate	6.17	6.16
	95% Credible Bounds	[5.95, 6.38]	[6.15, 6.18]
	Interval Width	0.43	0.03
$\widehat{R_g^2}$	Estimate	564.04	528.15
	95% Credible Bounds	[206.72, 920.42]	[493.51, 563.30]
	Interval Width	713.69	69.80

Table 5.1: Estimated intercept ($\widehat{\beta_0}$) and and squared radius of gyration ($\widehat{R_g^2}$) of the log intensity curve for the complex `H2AH2B` arising from the hybrid Gibbs-DRAM MCMC procedure of both constant-error and heteroskedastic models. The credible bounds are calculated using the empirical quantiles of the appropriate linear transformations of the $\boldsymbol{\theta}$ parameter chain.

		Constant Errors	**Heteroskedastic**
$\widehat{\beta_0}$	Estimate	6.16	6.16
	95% Credible Bounds	[5.96, 6.37]	[6.15, 6.17]
	Interval Width	0.41	0.02
$\widehat{R_g^2}$	Estimate	556.12	532.85
	95% Credible Bounds	[234.67, 877.58]	[503.02, 562.67]
	Interval Width	642.92	59.65

Table 5.2: Estimated intercept ($\widehat{\beta_0}$) and and squared radius of gyration ($\widehat{R_g^2}$) of the log intensity curve for the complex `H2AH2B` arising from variational approximations of both constant-error and heteroskedastic models. The credible bounds are derived using the appropriate linear transformation of the $\boldsymbol{\theta}$ under a Gaussian variational distribution

Table 5.1 contains the estimates of β_0 and R_g^2 for the MCMC procedure and Table 5.2 contains the corresponding estimates from the variational approximation. The point estimates for both β_0 and R_g^2 in all cases are fairly close. Appropriately accounting for the heteroskedastic nature of the data leads to more-appropriate uncertainty estimates on the physical parameters. In this case that means tighter bounds. Figures 5.9 and 5.10 show the fits of the MCMC procedure for the `H2AH2B` data.

A curious result was observed when applying the heteroskedastic models to `H2AH2B` data. For some cases, differences in magnitude of the covariates associated with $\boldsymbol{\theta}_V$ could cause erratic behavior of the resulting estimates. To address this, a mild scaling factor of 10 was applied to all values s. The corresponding R_g^2 and β_0 estimates account for this scaling. Crainiceanu et al. (2007) discuss issues with the scale of the spatial covariate in the context of the spatially adaptive heteroskedastic model and it is likely that similar convergence issues can be seen here. Investigating the effect of the spatial covariate scale, particularly on the multivariate sampling mechanism of the MCMC procedure for $\boldsymbol{\theta}_V$, is of interest for future work.

5.3 Joint Mean-Variance Example

The final portion of our case study of SAXS data is to analyze the hierarchical model developed in Section 2.3 for jointly modeling radial mean and variance data with experimental data. Once again, we consider a sampled subset of a 7-second exposure of `H2AH2B` complex taken at 4 mg/ml. However, this time we include the radial variance along with the radial mean intensity data. Figure 5.11 displays the data. While the log radial variance data displays an overall smooth behavior, there are a subset of points that appear to be gross outliers that occur in the region $[0.10, 0.20]$. These points are displayed as "X" in Figure 5.11. The corresponding radial mean values do not appear to be outliers. This raises a curious question as to the nature of these data. Most likely, this is tied to some sort of

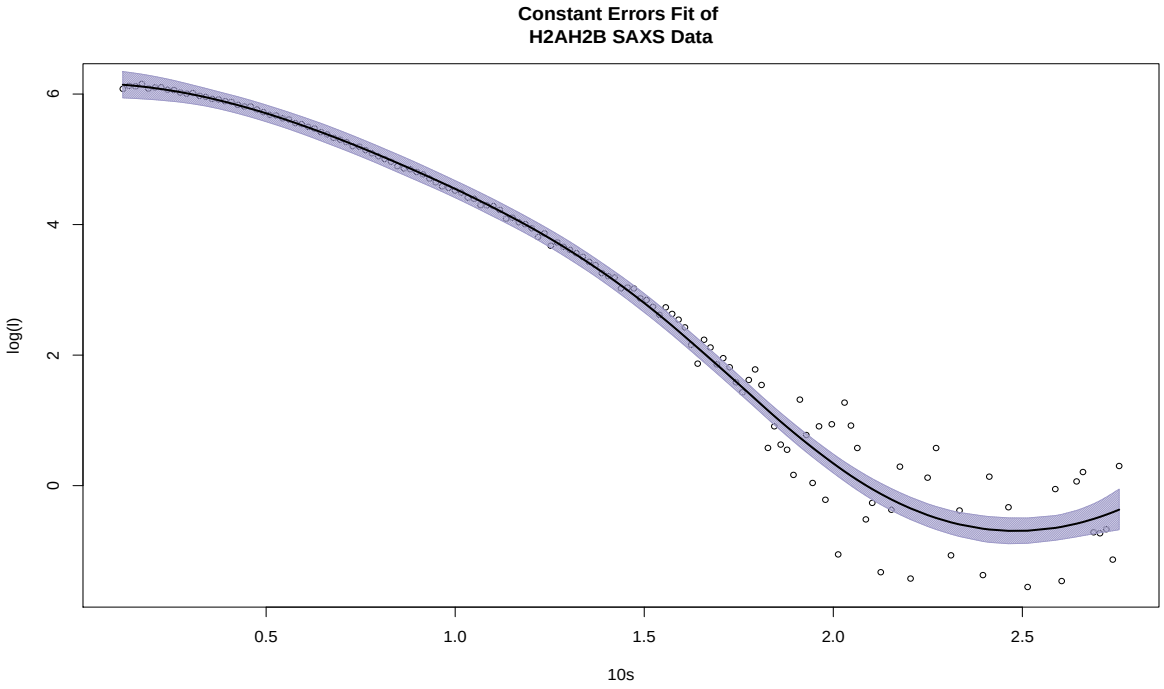


Figure 5.9: MCMC estimate via Gibbs sampling of a truncated quadratic penalized spline regression model with the linear fixed effect removed for the `H2AH2B` data. The procedure is run for 20000 iterations with a burn-in of 2000.

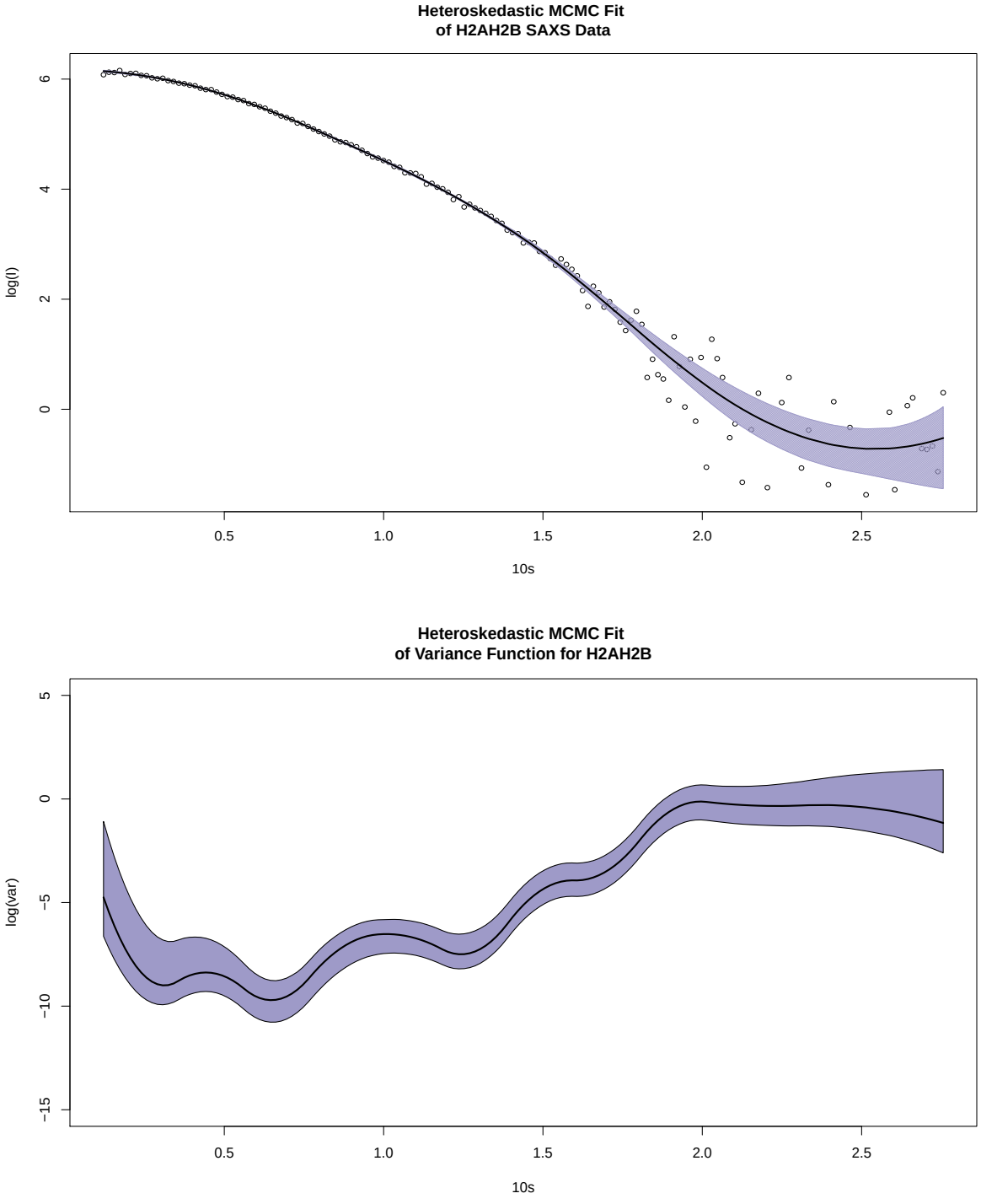


Figure 5.10: MCMC estimate via the hybrid Gibbs-DRAM procedure of a heteroskedastic truncated quadratic penalized spline regression model with the linear fixed effect removed for the `H2AH2B` data. The procedure is run for 40000 iterations with a burn-in of 4000.

detector failure in which dead pixels inflate those observed radial variance values. For now we remove these values from both responses.

To model both responses jointly, we consider the fixed effect radial variance model described in (40). For both the mean and model variance levels, a truncated quadratic basis is used, evaluated over $K = K_V = 10$ equally spaced quantiles from 0.1 to 0.99. For the radial variance level, a fixed effect model with covariates corresponding to an intercept, $\log(s)$, and the log model variance v is used. The MCMC procedure described in Section 3.3 is run for 75000 iterations with a 10% burn-in period. The resulting fits of the observed data are displayed in Figure 5.12 with the estimated log model variance function displayed in Figure 5.13

As with the simulated data example of Section 3.3.2, both the log radial mean intensity and radial variance data are fit quite well by the MCMC procedure. Surprisingly, the fixed effect model appears to do quite well for describing the log radial variance function. This is of note because the derived relationship of σ^2 and t^2 described in (17) is for the case where σ^2 describes the model variance of the intensity data, not the log intensity data. Table 5.3 shows the estimates for the fixed effect model used for the radial variance data. While the evidence suggests that η_2, the parameter for the log model variance covaraiate, is significantly different from 0, it is two orders of magnitude smaller than the $\log(s)$ parameter. This suggests that there may not be much auxiliary model variance information to be gleaned from the radial variance data.

	Estimate	2.5%	97.5%
η_0	0.756	0.692	0.823
η_1	-1.022	-1.074	-0.971
η_2	0.042	0.030	0.053
σ_u^2	0.013	0.010	0.017

Table 5.3: Parameter estimates for radial variance data of `H2AH2B` data.

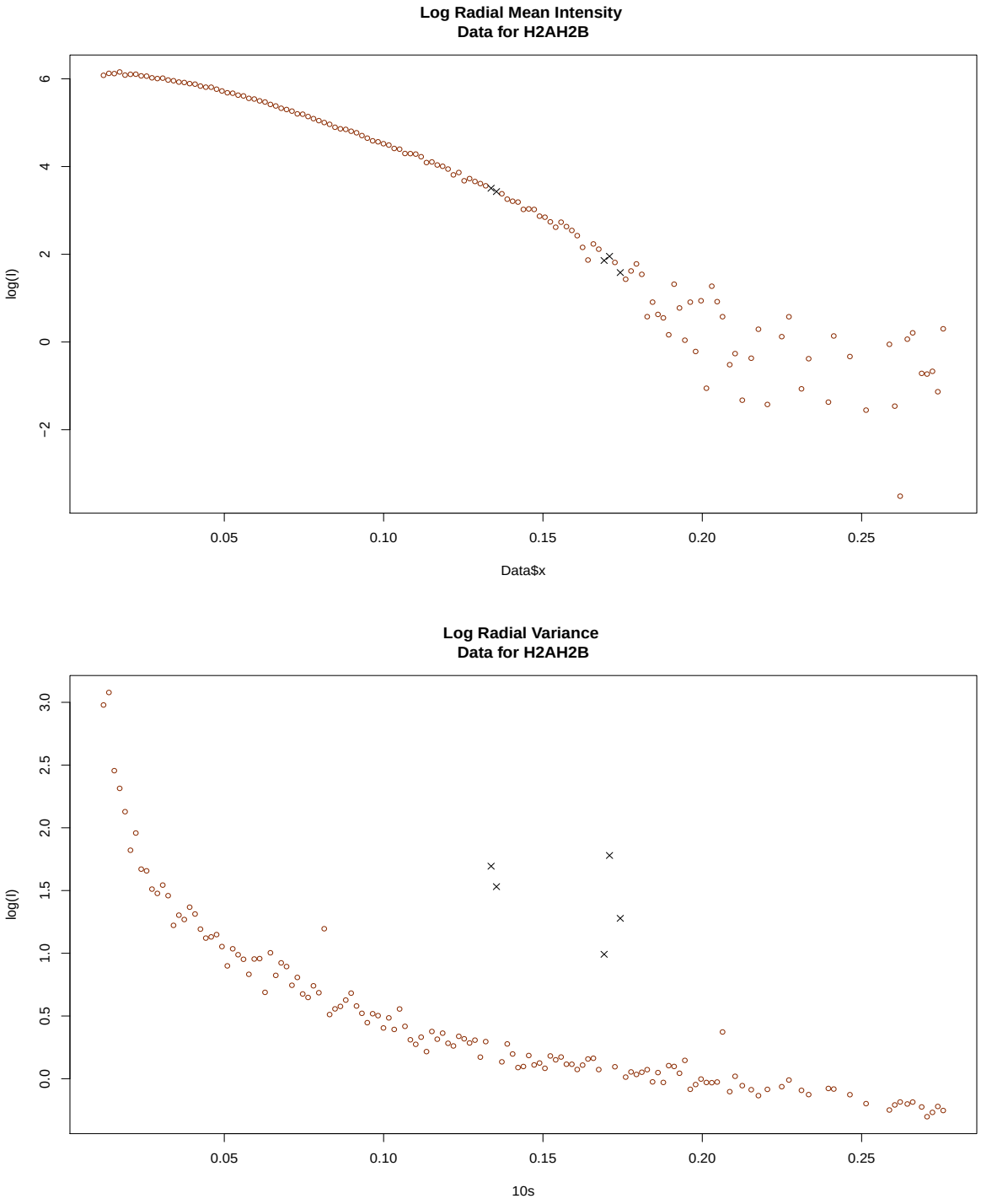


Figure 5.11: A sampled set of radial mean and variance data from a 7-second exposure of `H2AH2B` complex taken at 4 mg/ml. Data values denoted with an "X" represent outlier values on the log radial variance curves. Both the radial mean and radial variance responses associated with these values are removed for future analysis.

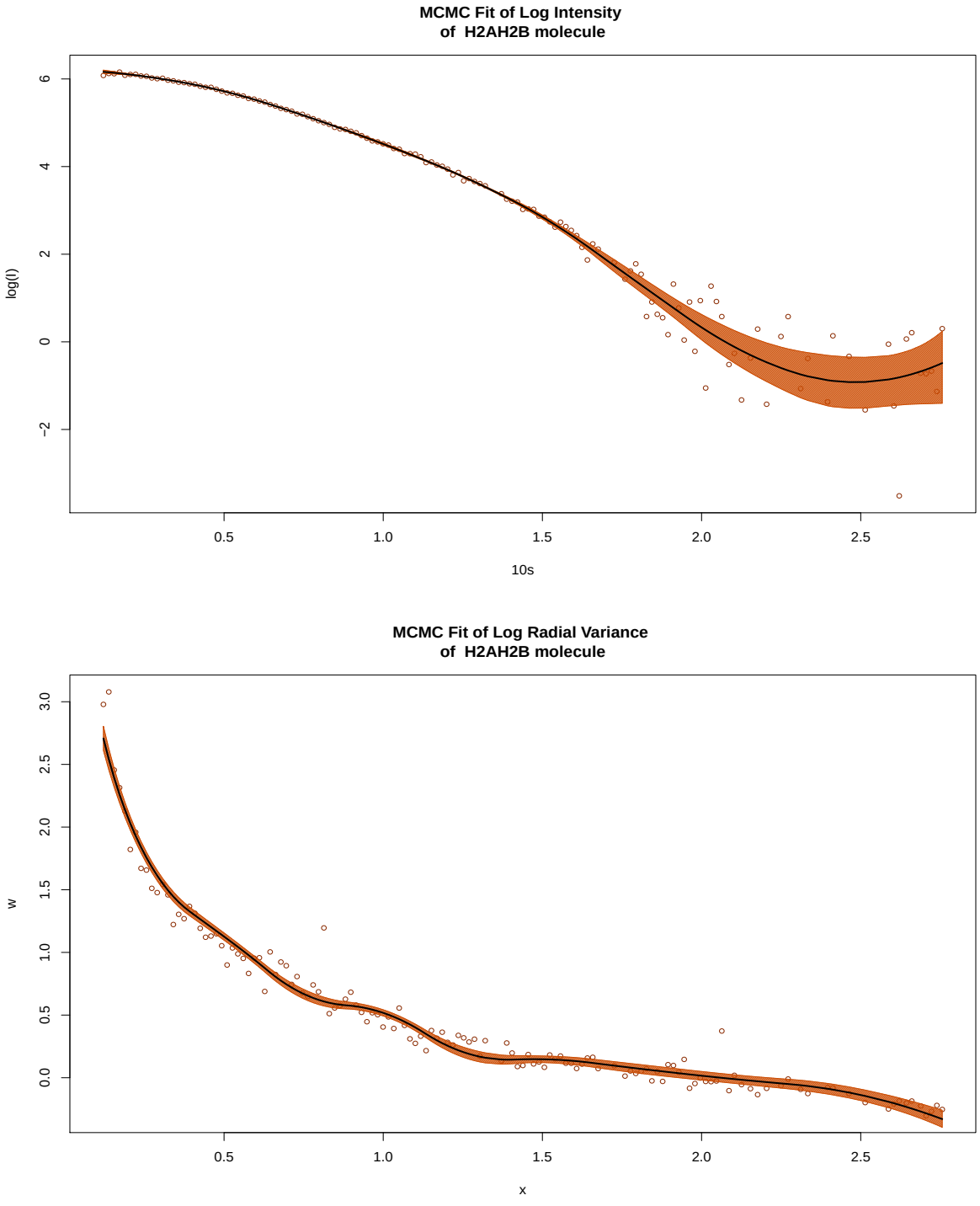


Figure 5.12: MCMC estimates of observed radial mean and variance data for a `H2AH2B` complex. The shaded regions correspond to 95% credible bounds for the estimated fits. The procedure was run for 75000 iterations with a 10% burn-in period.

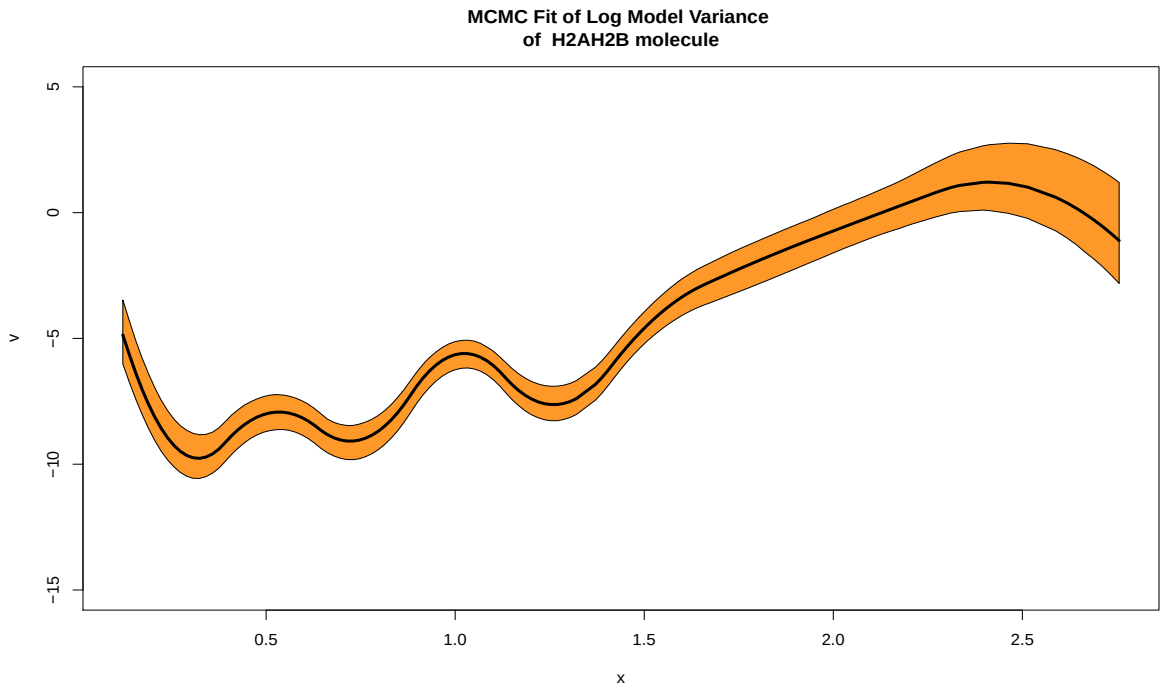


Figure 5.13: MCMC estimate of the log model variance function for the `H2AH2B` complex data. The shaded regions correspond to 95% credible bounds for the estimated fits. The procedure was run for 75000 iterations with stand-alonea 10% burn-in period.

The heteroskedastic nature of the radial mean data is reflected by the associated credible bounds in Figure 5.12. The estimated log model variance in Figure 5.13 is comparable to the estimate from the the purely heteroskedastic model in Figure 5.10. Figure 5.14 shows the post burn-in marginal trace plots of the $\boldsymbol{\theta}_V$ parameter chain. The overall acceptance rate of the DRAM step was approximately 13.6 percent. Some parameters in the chain appear to be converging rather slowly which could be addressed through further tuning of the DRAM procedure. Since these parameters represent coefficients of basis functions, it is unclear how slow convergence of the marginal chains affects the overall linear combination that fits the data.

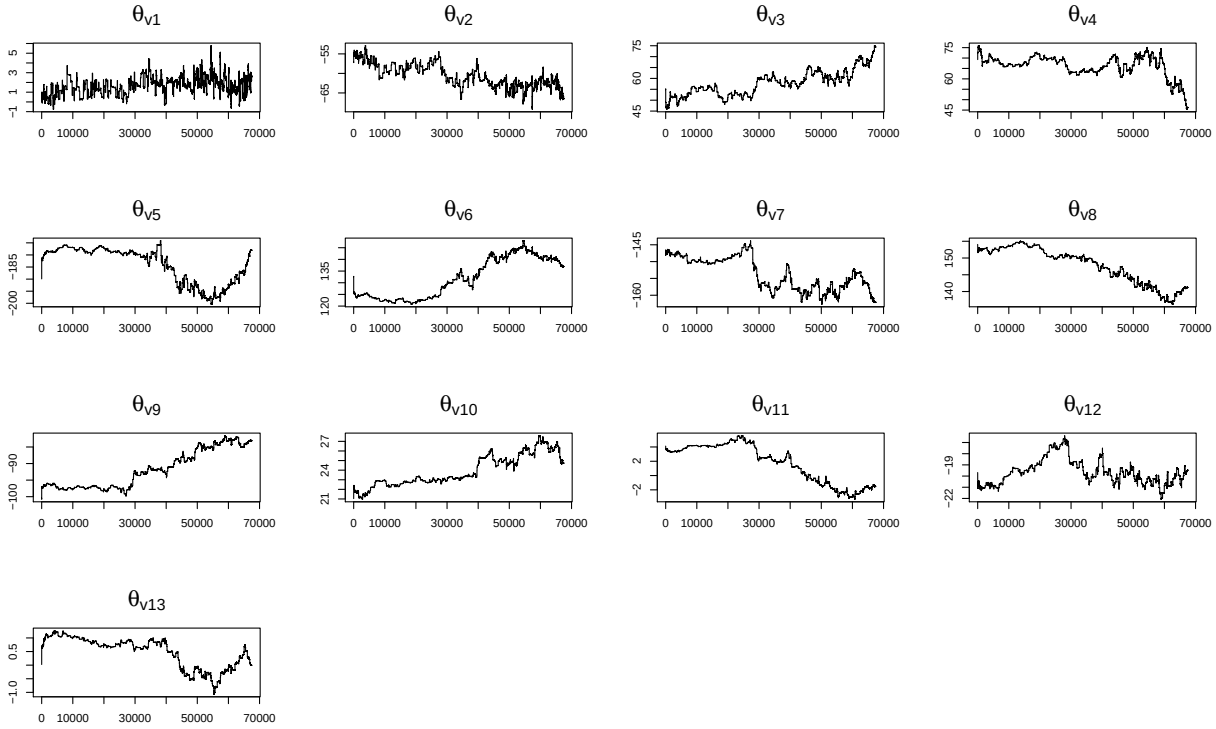


Figure 5.14: Post burn-in marginal trace plots of the $\boldsymbol{\theta}_V$ chain.

The inclusion of radial variance information in a model for SAXS data is an attempt to leverage as much information as possible to yield the most appropriate inference for both experimental effects and physical characteristics. The joint framework presented here is

able to fit both the simulation studies as well as experimental SAXS data accurately. After studying both types of examples, it is unclear as to the true effect of the radial variance data on the inference at the mean level, especially versus a stand-alone heteroskedastic model. However, radial variance data may serve a purpose for detecting experimental effects that are not obvious with mean level observation alone. Specifically, the radial variance data may be more sensitive to departures from the radial symmetry assumption that is made at the point of data collection. Building a cohesive framework that incorporates all available data for SAXS inference would allow for better understanding of both the experimental process and the physical molecule of interest.

5.4 Further Extensions for SAXS Data

The SAXS examples presented in this chapter have been relatively simple, considering no more than one or two exposures at a time. In these cases, the illustrative examples for all the models presented here can appear to be somewhat textbook. It is common for SAXS data to be more complicated, with exposures taken over multiple combinations of exposures times, concentrations, and other experimental conditions. Also, we have limited ourselves to presenting well understood, localized structural parameter estimates in the form of $\log(I(0))$ and R_g^2. Higher resolution structural information is a sought after goal of scientists and determining the relationship between shape parameters and model parameters quickly becomes challenging. Finally, as multiple exposures are considered, including radial variance data can provide an additional for avenue for the detection of "odd" data behavior such as sensor malfunction. Data arising from SAXS experiments is a rich source for future questions of both statistical and scientific interest.

CHAPTER 6

CONCLUSION

The prevailing theme of this work has been to study statistical problems arising in the modeling of data produced by small angle X-ray scattering (SAXS) experiments. SAXS provides a method for inferring low-resolution structural information for a wide class of complex and scientifically interesting macromolecules. Often times, the nature of these molecules make them ill-suited for analysis under competing experimental methodologies. However, numerous statistical challenges arise from SAXS data. In this work, we presented three Bayesian frameworks of increasing complexity for the fitting of semiparametric models to SAXS data. While motivated by problems in biochemistry, these methods are general and can be applied to many problems of similar nature. Also presented are fast, variational approximation methods for reducing the computational cost of these methods. Often times approximate inference is sufficient for initial analysis or is the only way to deal with large data issues in a feasible manner. Novel variational approximations for semiparametric regression in the presence of heteroskedastic errors as well as spatially adaptive semiparametric regression are described here. Finally, we detail how the methods presented in this dissertation can be used to investigate the effect of experimental conditions on SAXS data quality as well as estimate physical characteristics. Successfully implementing flexible and rigorous modeling methodologies that can be applied to SAXS data is of continuing interests to both statisticians and experimental scientists in the hopes of better understanding the structure of complex macromolecules.

Future research related to the work presented here will focus on three areas. The first area will be solving the computational problems associated with the heteroskedastic semipara-

metric regression and joint mean-variance models, primarily the slow convergence of the $\boldsymbol{\theta}_V$ chain. The second area will be continued investigation on the usage of the variational approximations presented here for semiparametric regression models under complex variance structures. This will include extending the work presented here to consider the spatially adaptive heteroskedastic regression model of Crainiceanu et al. (2007). Finally, we will continue exploring the practical applications of our methodologies for SAXS data, including the estimation of more informative physical parameters such as maximum linear dimension and the pairwise distance function of the molecule of interest. The research presented here provides a solid foundation for addressing all three of these areas of interest.

REFERENCES

Attias, H. (2000). A Variational Bayesian Framework for Graphical Models. *Advances in Neural Information Processing Systems*, 12(1-2):209–215.

Baladandayuthapani, V., Mallick, B. K., and Carroll, R. J. (2005). Spatially Adaptive Bayesian Penalized Regression Splines (P-splines). *Journal of Computational and Graphical Statistics*, 14(2):378–394.

Barreto, H. and Howland, F. (2005). *Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel*. Cambridge University Press.

Bergfors, T. M. (1999). *Protein Crystallization: Techniques, Strategies, and Tips: A Laboratory Manual*. International University Line, La Jolla.

Breidt, F. Jay, E., Andrea, and van der Woerd, M. (2012). Autocovariance Structures for Radial Averages in Small-Angle X-Ray Scattering Experiments. *Journal of Time Series Analysis*, 33(5):704–717.

Carroll, R. J. and Ruppert, D. (1982a). A Comparison Between Maximum Likelihood and Generalized Least Squares in a Heteroscedastic Linear Model. *Journal of the American Statistical Association*, 77(380):878–882.

Carroll, R. J. and Ruppert, D. (1982b). Robust Estimation in Heteroscedastic Linear Models. *The Annals of Statistics*, pages 429–441.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*, volume 30. CRC Press.

Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Joshi, A., and Goodner, B. (2007). Spatially Adaptive Bayesian Penalized Splines with Heteroscedastic Errors. *Journal of Computational and Graphical Statistics*, 16(2):265–288.

Damlen, P., Wakefield, J., and Walker, S. (1999). Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxillary Variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344.

Davidian, M. and Carroll, R. J. (1987). Variance Function Estimation. *Journal of the American Statistical Association*, 82(400):1079–1091.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2):89–102.

Faes, C., Ormerod, J., and Wand, M. (2011). Variational Bayesian Inference for Parametric and Nonparametric Regression with Missing Data. *Journal of the American Statistical Association*, 106(495):959–971.

Fischer, H., Oliveira Neto, M. d., Napolitano, H., Polikarpov, I., and Craievich, A. (2009). Determination of the Molecular Weight of Proteins in Solution from a Single Small-Angle X-Ray Scattering Measurement on a Relative Scale. *Journal of Applied Crystallography*, 43(1):101–109.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

Glatter, O. and Kratky, O. (1982). *Small Angle X-Ray Scattering*, volume 102. Academic press London.

Guinier, A. (1939). La Diffraction des Rayons X aux Très Petits Angles: Application a l'Étude de Phénoménes Ultramicrosopiques. *Annales de Physique*, 12:161–237.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). An Adaptive Metropolis Algorithm. *Bernoulli*, 7(2):223–242.

Isenberg, I. (1979). Histones. *Annual Review of Biochemistry*, 48:159–191.

Kammann, E. and Wand, M. P. (2003). Geoadditive Models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):1–18.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Minka, T., Winn, J., Guiver, J., and Knowles, D. (2010). `Infer.NET 2.4`. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

Minka, T. P. (2001). Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

Nott, D. J., Tran, M.-N., and Leng, C. (2012). Variational Approximation for Heteroscedastic Linear Models and Matching Pursuit Algorithms. *Statistics and Computing*, 22(2):497–512.

Opsomer, J. D., Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1999). Kriging with Nonparametric Variance Function Estimation. *Biometrics*, 55(3):704–710.

Ormerod, J. and Wand, M. (2010). Explaining Variational Approximations. *The American Statistician*, 64(2):140–153.

Pham, T. H., Ormerod, J. T., and Wand, M. (2013). Mean Field Variational Bayesian Inference for Nonparametric Regression with Measurement Error. *Computational Statistics & Data Analysis*, 68:375–387.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs Sampler: the Gibbs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, 87(419):861–868.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian Inference for Latant Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*, volume 12. Cambridge University Press.

Titterington, D. (2004). Bayesian Methods for Neural Networks and Related Models. *Statistical Science*, 19(1):128–139.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Wand, M. and Ormerod, J. (2008). On Semiparametric Regression with O'Sullivan Penalized Splines. *Australian & New Zealand Journal of Statistics*, 50(2):179–198.

Wang, C. and Blei, D. M. (2012). Variational Inference in Nonconjugate Models. *arXiv preprint arXiv:1209.4360*.

Wang, S. S. and Wand, M. P. (2011). Using `Infer.NET` for Statistical Analyses. Technical report, Centre for Statistical and Survey Methodology at University of Wollongong.

APPENDICES

A Derivation of q for Heteroskedastic Semiparametric Regression

The following section contains the derivation of the q-densities described in (76).

Derivation of $q_{\boldsymbol{\theta}}$

Using the form of $p(\boldsymbol{\theta} \mid \cdot)$ from (29), we can derive q_θ:

$$\begin{aligned}
q_\theta &\propto \exp\left[E_{-\theta}\left[\log p(\theta|\cdot)\right]\right] \\
&\propto \exp\left[E_{-\theta}\left[-\frac{1}{2}\left(\theta - \mathbf{M}\mathbf{C}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y}\right)^T \mathbf{M}^{-\mathbf{1}}\left(\theta - \mathbf{M}\mathbf{C}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y}\right)\right]\right] \\
&\propto \exp\left[E_{-\theta}\left[-\frac{1}{2}\left(\theta - \mathbf{M}\mathbf{C}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y}\right)^T \mathbf{M}^{-\mathbf{1}}\left(\theta - \mathbf{M}\mathbf{C}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y}\right)\right]\right] \\
&\propto \exp\left[-\frac{1}{2}E_{-\theta}\left[\theta^{\mathbf{T}}\mathbf{M}^{-\mathbf{1}}\theta - \mathbf{y}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{C}\theta - \theta^{\mathbf{T}}\mathbf{C}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y}\right]\right].
\end{aligned}$$

The quadratic term is

$$\begin{aligned}
E_{-\theta}\left[\theta^{\mathbf{T}}\mathbf{M}^{-\mathbf{1}}\theta\right] &= E_{-\theta}\left[\theta^{\mathbf{T}}\left(\boldsymbol{\Sigma}_\theta^{-\mathbf{1}} + \mathbf{C}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{C}\right)\theta\right] \\
&= \beta^{\mathbf{T}}\beta\frac{1}{\sigma_\beta^2} + \mathbf{b}^{\mathbf{T}}\mathbf{b}\underbrace{E_{-\theta}\left[\frac{1}{\sigma_b^2}\right]}_{\mu_{q(1/\sigma_b^2)}} + E_{-\theta}\left[\theta^{\mathbf{T}}\mathbf{C}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{C}\theta\right] \\
&= \beta^{\mathbf{T}}\beta\frac{1}{\sigma_\beta^2} + \mathbf{b}^{\mathbf{T}}\mathbf{b}\mu_{q(1/\sigma_b^2)} + \sum_{i=1}^{N}(C_i^T\theta)^2\underbrace{E_{-\theta}\left[\exp\left\{-C_{v_i}\theta_{\mathbf{V}}\right\}\right]}_{\gamma_i} \\
&= \theta^{\mathbf{T}}\left(\text{blockdiag}(\frac{1}{\sigma_\beta^2}\mathcal{I}_p, \mu_{q(1/\sigma_b^2)}\mathcal{I}_K) + \mathbf{C}^{\mathbf{T}}\boldsymbol{\Gamma}\mathbf{C}\right)\theta.
\end{aligned}$$

Define $\mathbf{\Sigma}_{q(\theta)} = \left(\text{blockdiag}(\frac{1}{\sigma_\beta^2}\mathcal{I}_p, \mu_{q(1/\sigma_b^2)}\mathcal{I}_K) + \mathbf{C^T\Gamma C}\right)^{-1}$. Here $\mathbf{\Gamma} = \text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_N)$. In this context $\gamma_i = E_{-\theta}\left[\exp\{-\mathbf{C_{V_i}}^T\theta_\mathbf{V}\}\right]$ corresponds to the moment generating function of q_{θ_V} evaluated at the vector $-\mathbf{C_{V_i}}$.

The expected value of the terms linear in θ is

$$\begin{aligned} E_{-\theta}\left[\mathbf{y^T\Sigma^{-1}C}\theta\right] &= E_{-\theta}\left[\sum_{i=1}^{N} y_i(\mathbf{C_i}^T\theta)\frac{1}{\sigma_i^2}\right] \\ &= \sum_{i=1}^{N} y_i(\mathbf{C_i}^T\theta)E_{-\theta_i}\left[-\mathbf{C_{V_i}}^T\theta_\mathbf{V}\right] \\ &= \mathbf{y^T\Gamma C}\theta. \end{aligned}$$

By symmetry, we also have

$$E_{-\theta}\left[\theta^\mathbf{T}\mathbf{C^T\Sigma^{-1}y}\right] = \theta^\mathbf{T}\mathbf{C^T\Gamma y}.$$

Using these results and completing the square, we arrive at the form of the optimal q_θ:

$$\begin{aligned} q_\theta &\propto \exp\left\{-\frac{1}{2}\left(\theta - \mathbf{\Sigma}_{q(\boldsymbol{\theta})}\mathbf{C}^T\mathbf{\Gamma y}\right)^T \mathbf{\Sigma_{q(\theta)}}^{-1}\left(\theta - \mathbf{\Sigma_{q(\theta)}C^T\Gamma y}\right)\right\} \\ \theta &\overset{q}{\sim} \mathcal{N}\left(\underbrace{\mathbf{\Sigma_{q(\theta)}C^T\Gamma y}}_{\mu_{q(\theta)}}, \mathbf{\Sigma_{q(\theta)}}\right). \end{aligned}$$

Derivation of $q_{\sigma_b^2}$

The variational density for σ_b^2 is

$$\begin{aligned} q_{\sigma_b^2} &\propto \exp\left[E_{-\sigma_b^2}\left[\log(p(\sigma_b^2 \mid \cdot)\right]\right] \\ &\propto \exp\left[E_{-\sigma_b^2}\left[-\left(A_b + \frac{K}{2}\right)\log(\sigma_b^2) - \frac{1}{\sigma_b^2}\left(B_b + \|\mathbf{b}\|^2\right)\right]\right] \\ &\propto \left(\sigma_b^2\right)^{-A_b - K/2}\exp\left[-\frac{1}{\sigma_b^2}\left[B_b + E_{-\sigma_b^2}[\|\mathbf{b}\|^2]\right]\right] \end{aligned}$$

$$\propto \left(\sigma_b^2\right)^{-A_b - K/2} \exp\left[-\frac{1}{\sigma_b^2} \left\{ B_b + \| \underbrace{E_{-\sigma_b^2}[\mathbf{b}]}_{\mu_{q(b)}} \|^2 + \text{trace}(\underbrace{\text{Var}_{-\sigma_b^2}(\mathbf{b})}_{\Sigma_{q(b)}}) \right\} \right]$$

$$\sigma_b^2 \overset{q}{\sim} \mathcal{IG}\left(A_b + \frac{K_V}{2}, \underbrace{B_b + \frac{1}{2}\left(\|\mu_{\mathbf{q(b)}}\|^2 + \text{trace}(\mathbf{\Sigma_{q(b)}}) \right)}_{B_{q(\sigma_b^2)}} \right).$$

Here $\mu_{q(\mathbf{b})}$ and $\mathbf{\Sigma}_{q(\mathbf{b})}$ refer to the portion of $\mu_{q(\theta)}$ and $\mathbf{\Sigma}_{q(\theta)}$ associated with the random effects parameters. Given the inverse Gamma distribution of $q_{\sigma_b^2}$

$$E_q\left[\frac{1}{\sigma_b^2}\right] = \frac{A_b + K/2}{B_{q(\sigma_b^2)}}.$$

Derivation of $q_{\sigma_c^2}$

Given the similar structure between $p(\sigma_b^2 \mid \cdot)$ and $p(\sigma_c^2 \mid \cdot)$, the derivation of $q_{\sigma_c^2}$ is the same.

$$q_{\sigma_c^2} \propto \left(\sigma_c^2\right)^{-A_c - K_V/2} \exp\left\{ -\frac{1}{\sigma_c^2} \left[B_c + \| \underbrace{E_{-\sigma_c^2}[\mathbf{c}]}_{\mu_{q(c)}} \|^2 + \text{trace}(\underbrace{\text{Var}_{-\sigma_b^2}(\mathbf{c})}_{\Sigma_{q(c)}}) \right] \right\}$$

$$\sigma_c^2 \overset{q}{\sim} \mathcal{IG}\left(A_c + \frac{K_V}{2}, \underbrace{B_c + \frac{1}{2}\left(\|\mu_{\mathbf{q(c)}}\|^2 + \text{trace}(\mathbf{\Sigma_{q(c)}}) \right)}_{B_{q(\sigma_c^2)}} \right)$$

As before, $\mu_{q(\mathbf{c})}$ and $\mathbf{\Sigma}_{q(\mathbf{c})}$ refer to the portion of $\mu_{q(\theta_\mathbf{V})}$ and $\mathbf{\Sigma}_{q(\theta_\mathbf{V})}$ associated with the random effects parameters. As with $1/\sigma_b^2$,

$$E_q\left[\frac{1}{\sigma_c^2}\right] = \frac{A_c + K_V/2}{B_{q(\sigma_c^2)}}.$$

Derivation of $q_{\boldsymbol{\theta}_V}$

As highlighted in (29), the posterior conditional $p(\theta_{\mathbf{V}} \mid \cdot)$ does not take known form. Using (57), the optimal q-density for $\theta_{\mathbf{V}}$ is

$$
\begin{aligned}
q_{\theta_V} &\propto \exp\left[E_{-\theta_V}\left\{\log p(\theta_{\mathbf{V}} \mid \cdot\right\}\right] \\
&\propto \exp\left[E_{-\theta_V}\left\{-\frac{1}{2}\left(\sum_{i=1}^{N} C_{V_i}\theta_{\mathbf{V}} + \sum_{i=1}^{N}(y_i - \mathbf{C_i}^T\theta)^2\exp(-\mathbf{C_{V_i}}^T\theta_{\mathbf{V}}) + \theta_{\mathbf{V}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\theta_{\mathbf{V}}}^{-\mathbf{1}}\theta_{\mathbf{V}}\right)\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{N}\mathbf{C_{V_i}}^T\theta_{\mathbf{V}} + \sum_{i=1}^{N}E_{-\theta_V}\left[(y_i - \mathbf{C_i}^T\theta)^2\right]\exp(-\mathbf{C_{V_i}}^T\theta_{\mathbf{V}}) + E_{-\theta_V}\left[\theta_{\mathbf{V}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\theta_{\mathbf{V}}}^{-\mathbf{1}}\theta_{\mathbf{V}}\right]\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{N}\mathbf{C_{V_i}}^T\theta_{\mathbf{V}} + \sum_{i=1}^{N}\left((y_i - \mathbf{C_i}^T\mu_{q(\theta)})^2 + \mathbf{C_i}^T\boldsymbol{\Sigma}_{\mathbf{q}(\theta)}\mathbf{C_i}\right)\exp(-\mathbf{C_{V_i}}^T\theta_{\mathbf{V}})\right.\right. \\
&\quad \left.\left. + \theta_{\mathbf{V}}{}^T\text{blockdiag}\left(\frac{1}{\sigma_\eta^2}\mathcal{I}_r, \mu_{q(1/\sigma_c^2)}\mathcal{I}_{K_V}\right)\theta_{\mathbf{V}}\right\}\right].
\end{aligned}
$$

B Derivation of K-L Lower Bound for Heteroskedastic Semiparametric Regression

Let Γ represent the gamma function arising from the density of the inverse gamma distribution. Recall that the density $q(\boldsymbol{\psi})$ is assumed to have the product density structure described in Section 4.3. Then

$$
\begin{aligned}
\log \underline{p}(\mathbf{y}; q) &= \int_{\Psi} q(\boldsymbol{\psi})\left[\log p(\mathbf{y}, \boldsymbol{\psi}) - \log q(\boldsymbol{\psi})\right] d\boldsymbol{\psi} \\
&= \int_{\Psi} q(\boldsymbol{\psi})\left[\log p(\mathbf{y} \mid \boldsymbol{\psi}) + \log p(\boldsymbol{\psi}) - \log q(\boldsymbol{\psi})\right] d\boldsymbol{\psi} \\
&= \int_{\Psi} q(\boldsymbol{\psi})\left[\log p(\mathbf{y} \mid \boldsymbol{\psi}) + \log p(\boldsymbol{\theta}) + \log p(\sigma_b^2) + \log p(\boldsymbol{\theta}_V) + \log p(\sigma_c^2) - \log q(\boldsymbol{\psi})\right] d\boldsymbol{\psi} \\
&= -\frac{1}{2}\log(2\pi)N - \frac{p}{2}\log(\sigma_\beta^2) - \frac{r}{2}\log(\sigma_\delta^2) + A_b\log(B_b) - \log(\Gamma(A_b))
\end{aligned}
$$

$$
\begin{aligned}
&+ A_c \log(B_c) - \log(\Gamma(A_c)) + \frac{1}{2}\log(|\mathbf{\Sigma}_{q(\boldsymbol{\theta})}|) + \frac{1}{2}\log(|\mathbf{\Sigma}_{q(\boldsymbol{\theta}_V)}|) \\
&- \left(A_b + \frac{K}{2}\right)\log(B_q(\sigma_b^2)) - \left(A_c + \frac{K_V}{2}\right)\log(B_{q(\sigma_c^2)}) \\
&+ \log\left(\Gamma\left(A_C + \frac{K_V}{2}\right)\right) + \log\left(\Gamma\left(A_b + \frac{K}{2}\right)\right) \\
&+ \frac{1}{2\sigma_\beta^2}\left(\|\mu_{q(\beta)}\|^2 + \text{trace}(\mathbf{\Sigma}_{q(\beta)})\right) + \frac{1}{2\sigma_\eta^2}\left(\|\mu_{q(\delta)}\|^2 + \text{trace}(\mathbf{\Sigma}_{q(\eta)})\right) \\
&- \frac{1}{2}\sum_{i=1}^{N}\left[\left(Y_i - \mathbf{C}_i^T\mu_{q(\boldsymbol{\theta})}\right)^2 + \mathbf{C}_i^T\mathbf{\Sigma}_{q(\boldsymbol{\theta})}\mathbf{C}_i\right]\exp\left\{-\mathbf{C}_{V_i}^T\mu_{q(\boldsymbol{\theta})} + \frac{1}{2}\mathbf{C}_{V_i}^T\Sigma_{q(\boldsymbol{\theta})}\mathbf{C}_{V_i}\right\} \\
&+ \frac{p+K}{2} + \frac{r+K_V}{2}.
\end{aligned}
$$